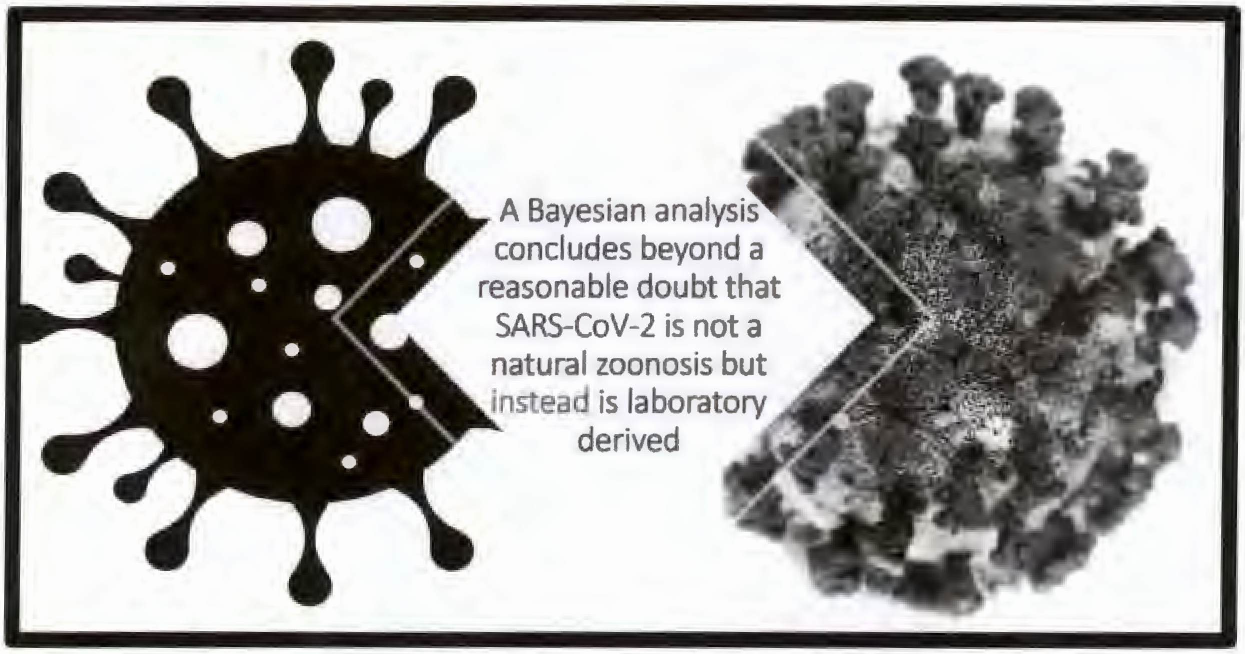


A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

Wuhan Institute of Virology analysis of bronchial lavage specimens from ICU patients at Wuhan Jinyintan Hospital in December 2019 contain both SARS-CoV-2 and adenovirus vaccine sequences, consistent with a vaccine challenge trial

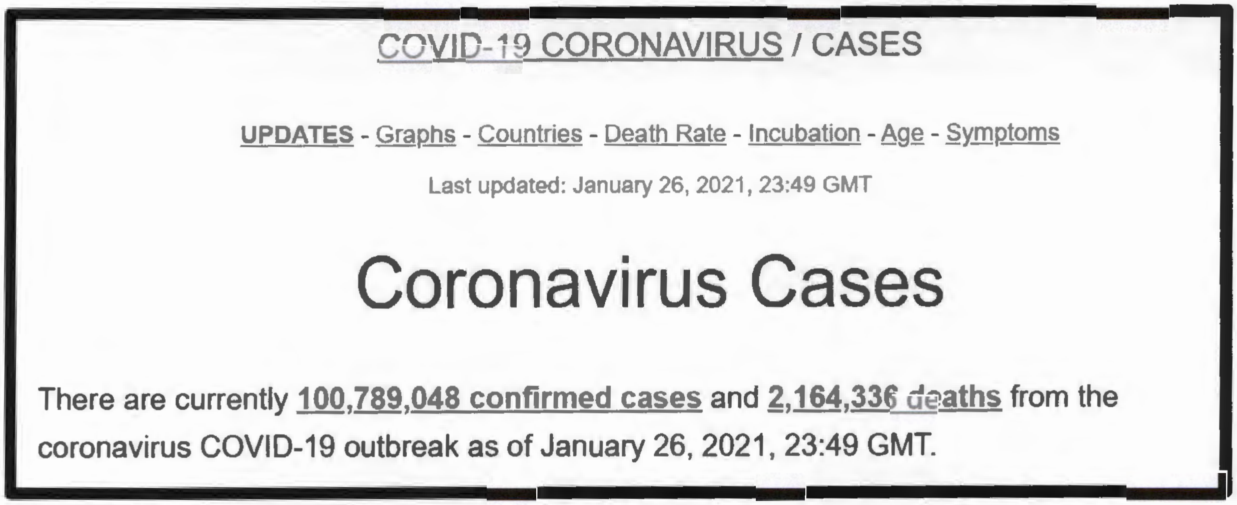
By

Steven Carl Quay, MD, PhD

www.DrQuay.com

Steven@DrQuay.com

Dedication. This work is dedicated to the men, women, and children who were infected with SARS-CoV-2 over the last year. It is my hope that this work becomes part of the body of evidence to help inform the public about gain-of-function pathogen research and that a renewed debate can be had about the benefits and risks of this research in the context of world health.



COVID-19 CORONAVIRUS / CASES

UPDATES - Graphs - Countries - Death Rate - Incubation - Age - Symptoms

Last updated: January 26, 2021, 23:49 GMT

Coronavirus Cases

There are currently 100,789,048 confirmed cases and 2,164,336 deaths from the coronavirus COVID-19 outbreak as of January 26, 2021, 23:49 GMT.

Source: https://www.worldometers.info/coronavirus/

Acknowledgements. Despite having collaborated over many decades on numerous scientific projects, research during 2020 into COVID-19, SARS-CoV-2, and therapeutic approaches has been a unique experience. With lockdowns and international travel bans, all collaborative work has been virtual. With an apparent bias surrounding investigation into the origin of SARS-CoV-2, ad hoc groups of Citizen-Scientists, often anonymous, have worked together via email, videoconference, micro-blogging, and social messaging networks to advance our understanding of this horrific pandemic.

I want to thank a Twitter group called #DRASTIC for many useful discussions that found their way into this document. Dr. Martin Lee, Ph.D., Adjunct Professor of Statistics at UCLA provided statistical support throughout this work. H. Lawrence Remmel provided input on the adenovirus vaccine as a dual target vaccine. I want to thank D.A. for originally suggesting performing a Bayesian analysis on the work I had done on SARS-CoV-2 and for his facilitation of the review of this work by a diverse group of scientists and policy makers.

In all cases, however, this is my own work product.

A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

Wuhan Institute of Virology analysis of lavage specimens from ICU patients at Wuhan Jinyintan Hospital in December 2019 contain both SARS-CoV-2 and adenovirus vaccine sequences consistent with a vaccine challenge trial

Executive Summary. The one-year anniversary of the COVID-19 pandemic records 2.1 million deaths, over 100 million confirmed cases,[1] and trillions of dollars of economic damage. Although there is universal agreement that a coronavirus identified as Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 (abbreviated CoV-2 henceforth) causes the disease COVID-19, there is no understanding or consensus on the origin of the disease.

The Chinese government, WHO, media, and many academic virologists have stated with strong conviction that the coronavirus came from nature, either directly from bats or indirectly from bats through another species. Transmission of a virus from animals to humans is called a zoonosis.

A small but growing number of scientists have considered another hypothesis: that an ancestral bat coronavirus was collected in the wild, genetically manipulated in a laboratory to make it more infectious, training it to infect human cells, and ultimately released, probably by accident, in Wuhan, China. For most of 2020 this hypothesis was considered a crackpot idea, but in the last few weeks, more media attention has been given to the possibility that the Wuhan Institute of Virology, located near the Wuhan city center and with a population of over 11 million inhabitants, may have been the source of the field specimen collection effort, laboratory genetic manipulation, and subsequent leak. On January 15, 2021, the U.S. Department of State issued a statement requesting the WHO investigation of the origin of COVID-19 include specific assertions related to a laboratory origin of the pandemic.[2]

Given the strong sentiment in the scientific community in favor of a zoonosis and the massive effort undertaken by China to find the natural animal source, one can assume that any evidence in favor of a natural origin, no matter how trivial, would become widely disseminated and known. This provides a potential evidence bias within the scientific community in favor of a natural origin which isn't quantifiable but should be kept in mind.

This becomes especially important background when evidence that could support a laboratory origin has been directly provided by leading Chinese scientists themselves, like Dr. Zhengli Shi, head of coronavirus research at the Wuhan Institute of Virology and Gao Fu (George Fu Gao), Director of Chinese CDC; by the Chinese government, as well as by powerful and vocal, pro-natural origin scientists, like Dr. Peter Daszak, of the NYC-based NGO, EcoHealth Alliance.

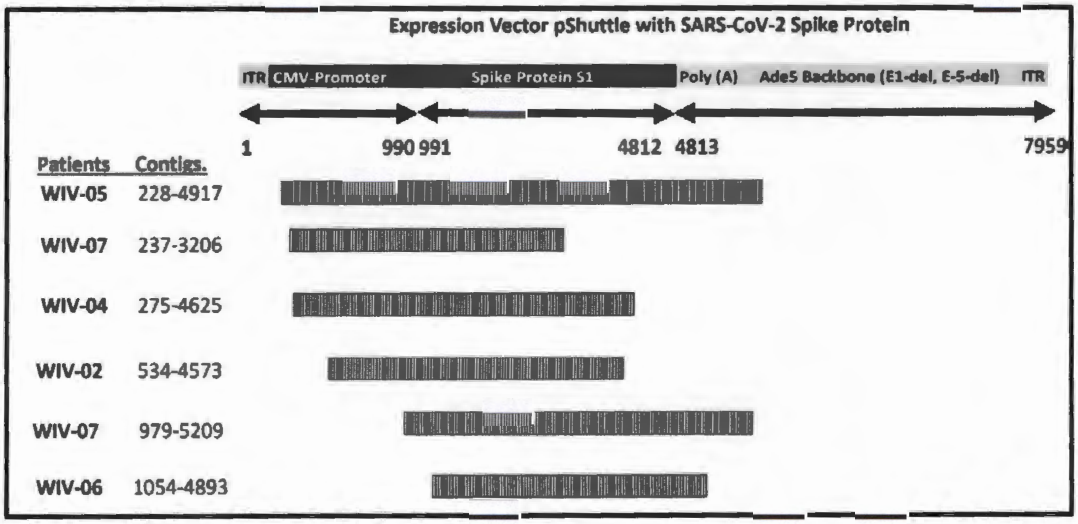
[1] https://www.worldometers.info/coronavirus/?
[2] https://www.state.gov/ensuring-a-transparent-thorough-investigation-of-covid-19s-origin/

This report uses Bayesian inference, a common statistical tool in which Bayes' theorem, a well-known statistical equation, is used to update the likelihood for a particular hypothesis as more evidence or information becomes available. It is widely used in the sciences and medicine and has begun to be used in the law.

The starting probability for origin of SARS-CoV-2 was set with the zoonotic or natural hypothesis at 98.8% likelihood with the laboratory origin hypothesis set at 1.2%. The initial state was biased as much as possible towards a zoonotic origin, with the starting point selected as the upper bounds of the 95% confidence interval for the mean and standard deviation of three independent estimates, including one by Daszak and colleagues. Each piece of new evidence for or against each hypothesis was then used to adjust the probabilities. If evidence favored a natural origin the math adjusts upward the probability of a natural origin, and so on.

The most significant evidence provided herein is the finding from RNA-Seq performed by the Wuhan Institute of Virology (WIV) of lavage patient samples collected on December 30, 2019.[3] These ICU patients were the subject of the seminal paper, entitled, "A pneumonia outbreak associated with a new coronavirus of probable bat origin," from Dr. Zhengli Shi and colleagues that first characterized SARS-CoV-2.[4] This author has confirmed that the RNA-Seq of all five patients contained SARS-CoV-2 sequences.

Surprisingly the specimens also contained the adenovirus "pShuttle" vector, developed by Chinese scientists in 2005 for SARS-CoV-1.[5] Two immunogens were identified, the Spike Protein gene of SARS-CoV-2 and the synthetic construct H7N9 HA gene.[6] Hundreds of perfectly homologous (150/150) raw reads suggest this is not an artefact. Reads that cross the vector-immunogen junction are identified. An example of the read contigs for CoV-2 is shown in this figure:



[3] The detailed evidence for the adenovirus vaccine sequences is given at the end of this document.
[4] https://www.nature.com/articles/s41586-020-2012-7
[5] https://www.ncbi.nlm.nih.gov/nuccore/AY862402.1
[6] https://www.ncbi.nlm.nih.gov/nuccore/KY199425.1/

While adenovirus is a common infection the wildtype viruses have low homology to the vaccine vector sequence, by design, to avoid rejection of the vaccine due to prior exposure to wildtype adenoviruses.

Two patients from the same hospital who had bronchial lavage on the same day but had their specimens sent to the Hubei CDC did not have adenovirus vaccine sequences.

Three explanations come to mind from this evidence:

1. **These represent sample preparation artifacts at the WIV, such as sample spillover on the sequencer.**
2. **These patients were admitted with an unknown infection, were not responding to the treatment protocols for a infection of unknown origin, and they were vaccinated with an experimental vaccine in a desperate but compassionate therapeutic "Hail Mary."**
3. **A clinical trial of a combination influenza/SARS-CoV-2 vaccine was being conducted and an accidental release into Wuhan occurred.**

Only WIV scientists and Chinese authorities can answer these questions. Until the evidence of the adenovirus sequences has been confirmed by other scientists, this author will not include this evidence in the Bayesian analysis.

Obviously if a vaccine containing the Spike Protein of SARS-CoV-2 was being administered to patients in Wuhan in December 2019 the question of laboratory origin is a settled matter.

The remaining analysis is being conducted without the adenovirus vaccine evidence unless and until it is corroborated. The outcome of this report is the conclusion that the probability of a laboratory origin for CoV-2 is 99.8% with a corresponding probability of a zoonotic origin of 0.2%. This exceeds most academic law school discussions of how to quantify 'beyond a reasonable doubt,' the threshold for finding guilt in a criminal case. The report contains the detailed analysis and quantitative basis for the statistics and conclusion. It should be noted that because of the commutative property of the collected adjustments to the probabilities, the order in which they are used in the overall calculation is immaterial and the same end likelihoods will be reached regardless of the order of input.

The following Text-Table summarizes the evidence examined and the changes in probabilities:

Evidence	Zoonotic Origin	Laboratory Origin
Initial State	98.8%	1.2%
International committees to determine CoV-2 origin may not be impartial	98.8%	1.2%
Three key zoonotic papers: pros and cons	98.8%	1.2%
SARS-like infections among employees of the Wuhan Institute of Virology in the fall of 2019 reported by US Government	98.8%	1.2%
Location of first cases near Wuhan Institute of Virology	95.1%	4.9%
Lack of evidence of seroconversion in Wuhan and Shanghai	80.9%	19.1%
Lack of posterior diversity	30.8%	69.2%
Opportunity: The Wuhan Institute of Virology has publicly disclosed that by 2017 it had developed the techniques to collect novel coronaviruses, systematically modify the receptor binding domain to improve binding or alter zoonotic tropism and transmission, insert a furin site to permit human cell infection, make chimera and synthetic viruses, perform experiments in humanized mice, and optimize the ORF8 gene to increase human cell death.	30.8%	69.2%
Lack of furin cleavage sites in any other sarbecovirus	4.7%	95.3%
Rare usage of -CGG- single codons & no CGG-CGG pairs	0.5%	99.5%
Routine use of CGG in laboratory codon optimization, including Daszak & Shi	0.2%	99.8%
Spike Protein receptor binding region (200 amino acids) optimized for humans	0.2%	99.8%
Whole genome analysis shows pre-adaption of CoV-2	0.2%	99.8%
The finding of CoV-2 in Barcelona wastewater in early 2019 was an artifact	0.2%	99.8%
Shi and the WHO comment early on that CoV-2 seemed to begin with a single patient	0.2%	99.8%
Mammalian biodiversity between Yunnan and Hubei is significantly different, limiting a potential common intermediate host	0.2%	99.8%
The ancestor of CoV-2 can only obtain a furin site from other subgenera viruses but recombination is limited/non-existent between subgenera	0.2%	99.8%
Canvas of 410 animals shows humans and primates are the best, bats are the worst, for ACE2-Spike Protein interaction	0.2%	99.8%
A government requested review of samples collected from a mineshaft may have caused the COVID-19 pandemic	0.2%	99.8%
The Hunan Seafood Market and farmed animals in Hubei province are not the source of CoV-2	0.2%	99.8%
Line 2 of the Wuhan Metro System is the likely conduit of the pandemic and is the closest subway line to the WIV	0.2%	99.8%
Feral and domestic cats are not the intermediate host	0.2%	99.8%
Extraordinary pre-adaption for the use of human tRNA is observed	0.2%	99.8%
Evidence of lax operations and disregard of laboratory safety protocols and regulations in China	0.2%	99.8%
Previous SARS-CoV-1 laboratory accidents	0.2%	99.8%
Shi and Daszak use Wuhan residents as negative control for zoonotic coronavirus exposure	0.2%	99.8%
RaTG13 could be CoV-2 precursor using the synthetic biology 'No See 'Em' technique	0.2%	99.8%
Location, location, location: Based on the distance between known SARS-CoV-1 laboratory-acquired infections and the hospital of admission of the infected personnel, the WIV is within the expected hospital catchment for a CoV-2 LAI	0.2%	99.8%

The summary which follows will simply be a review and discussion of the evidence in the context of the two hypotheses.

Zoonosis Hypothesis

A viral zoonosis has at least three elements, a host, a virus, and the human population. With some viruses there are often two hosts. One is a 'reservoir host' where the virus can live for years or even decades in a relatively stable relationship. The reservoir host is never decimated by the virus, and the virus is never burned out by the reservoir host, disappearing completely. For coronaviruses the reservoir host is always one or more bat species. If there is a reservoir host that some viruses that cannot jump directly into the human population, there is a need for an second host, an intermediate host. In this case the virus spends time jumping into the intermediate host, 'practicing' adaption through random mutation and Darwinian selection for fitness to reproduce, infect, and transmit in the intermediate host. This process is then repeated between the intermediate host and the human population. Alternatively, the virus can jump directly between the bat reservoir and humans, without the need for an intermediate host.

For two prior human coronavirus epidemics, an intermediate or proximate host was identified. For SARS-CoV-1 in 2003-4 it was the civet cat while for Middle Eastern Respiratory Syndrome (MERS) in 2012-4 it was the camel. In both of these human epidemics, the intermediate host was identified within four to ten months of the first clinically identified human infection. With CoV-2 we are at 12 months since the pandemic began and still waiting for evidence of, despite a much larger effort inside China to find an intermediate host. For both of these previous pandemics, a bat species reservoir host was also identified, but not in the case of SARS-CoV-2.[7]

Based on the genome sequence of CoV-2, Drs. Shi and Daszak have proposed that the reservoir host for CoV-2 is the intermediate horseshoe bat (*Rhinolophus affinis*), which is found in Yunnan Province. Yunnan Province is in southern, rural China and about 1900 km from the north central province of Hubei, where the 11 million people of Wuhan live. In the US this would be equivalent in distance, climate change, and human population density difference to going from the Everglades in Florida to Manhattan, in New York City. The intermediate horseshow bat isn't found at all in Hubei province, making a direct bat-to-human transmission improbable.[8] Experiments in three independent laboratories also demonstrate that CoV-2 has changed genetically so much that it can no longer infect any bat species cell culture tested. So, while the leading US coronavirus expert, Dr. Ralph Baric of The University of North Carolina suggested in early 2020 that CoV-2 may have jumped into the human population directly from bats without an intermediate host, this hypothesis seems to no longer be viable.

For the zoonosis hypothesis to be advanced, it is now necessary to find an intermediate host. In January 2020 a theory was proposed that CoV-2 arose in the Huanan Seafood Market, a traditional Chinese "wet market" where live animals are butchered and sold for food. The market theory was based on the observation that about 40% of early patients worked or shopped there. This was reminiscent of the wet market sources for civet cats infected with SARS-CoV-1 or the camel markets for the MERS coronavirus. The Chinese authorities closed the market on December 31, 2019 after performing extensive environmental sampling and sanitation.

But by May 2020 Dr. Gao Fu, Director of the Chinese CDC, announced that the market was not the source of CoV-2, as all of the animal specimens tested negative for CoV-2. And while SARS-CoV-1 was found in 100% of local farmed civets when tested, CoV-2 was different. In July 2020 Dr. Shi reported that extensive testing of farmed animals throughout Hubei Province failed to find CoV-2 in any animals.

For about six months, the pangolin, a scaly anteater, was suspected to be the intermediate host but finally Dr. Daszak reported that CoV-2 was not found in pangolins in the wild or from the (illegal) market trade.[9] Domestic and feral cats also were ruled out as a possible source. A

[7] I am distinguishing here the difference between SARS-CoV-2 being a descendent of a bat coronavirus (with 3.8% or 1100 nucleotide (nt) differences between them) and the finding of the immediate precursor of SARS-CoV-2 in a bat colony population somewhere in the wild, which usually is <100 nt differences.

[8] "We have done bat virus surveillance in Hubei Province for many years but have not found that bats in Wuhan or even the wider Hubei Province carry any coronaviruses that are closely related to SARS-CoV-2. I don't think the spillover from bats to humans occurred in Wuhan or in Hubei Province," said Dr. Shi. Science, July 2020

[9] https://link.springer.com/article/10.1007/s10393-020-01503-x

comprehensive computer-based screen of 410 different animals reported the remarkable finding that the best ACE2 receptor matches to CoV-2 were human and other primates (or primate cells in the laboratory), including the favorite laboratory coronavirus host, the VERO monkey cell culture, and that all bat species were the worst host. At the time of this writing, there is not even a working hypothesis for the species of an intermediate host.

A typical zoonosis has a number of characteristic properties that can allow identification of a zoonotic infection, even in the absence of identifying an intermediate host. None of these properties are found for CoV-2.

All zoonotic infections have in common the principle that when a virus in nature uses evolution to move from, for example, a bat host to a camel host and then to a human host, it is a hit and miss, slow process. After all, evolution is the result of random genetic changes, mutations, and then enrichment of the ones that are helpful by amplification during reproduction. With both SARS-CoV-1 and MERS, the coronavirus spent months and years jumping from the intermediate host into humans, not having all of the necessary mutations needed to be aggressive, grow, and then spread, but spending enough time in humans to cause an infection and leaving behind a corresponding immune response.

The hallmark evidence of this 'practice' in abortive host jumping is in stored, archived human blood specimens taken from before the epidemic, where one can find evidence of pre-epidemic, usually sub-clinical, community spread from the antibodies to the eventual epidemic virus. For SARS-CoV-1 and MERS, about 0.6% of people in the region where the epidemic began showed signs of an infection in archived blood. With CoV-2, this seroconversion, as it is called, has never been observed, including in 540 specimens collected from 'fever clinics' in Wuhan between October 2019 and January 2020, reported by the WHO. Because this is such a potent signal of a zoonosis, and because I believe that China has over 100,000 stored specimens from Wuhan taken in the fall of 2019, the lack of reports of seroconversion, the silence from China on this evidence, speaks volumes.

Another hallmark of a slow, natural zoonosis can be found in the virus. In SARS-CoV-1 and MERS, the coronavirus spent years in the intermediate host, passing back and forth among populations of hosts, the civets or camels, that were living in close proximity. During this time, they would accumulate a background of genetic mistakes, i.e., mutations- usually about one mistake every two weeks. When the final chip falls, and a mutation(s) happens allowing the jump into humans, the virus with that new mutation(s) also jumps around within the intermediate host population. The consequence of this latter behavior for a true zoonosis is that the genome sequences found in humans don't all descend from a single jump into a single human but show jumps from viruses that are only cousins of each other, not direct lineal descendants.

In a true zoonosis, the family tree of virus genome sequences doesn't pass back through the first patient but instead tracks all the way back to an ancestor months or years earlier. This is called posterior diversity, and it is an easy genetic test to perform. With CoV-2, every one of the more than 294,000 virus genomes sequenced can be traced back to the first genomic cluster and in the first patient in that cluster, a 39-year-old man who was seen at the People's Liberation Army

(PLA) Hospital about one mile from the Wuhan Institute of Virology. The CoV-2 pandemic has the phylogenetic signature of one pure virus sequence infecting one human, with human-to-human spread thereafter; there is just the one and only jump into the human population ever seen. This lack of posterior diversity has been alluded to by Dr. Shi, by the WHO, and by other prominent virologists; they just never take that critical piece of the evidence to the next the proper inference.

The virus in a true zoonosis also contains the signature record of the gradual changes and adaptions it made in the protein key, the Spike Protein, it uses to unlock human cells and cause infection. With SARS-CoV-1 the Spike Protein had fewer than one-third of all the changes it would later develop by the time it became an epidemic. With CoV-2 the Spike Protein was almost perfectly adapted to the human lock, using 99.5% of the best amino acids possible.

Since with CoV-2 we have no evidence from stored blood that it was quietly practicing on humans in the community of Wuhan, it is surprising that when it finds its first patient, it has perfected to 99.5% the spike protein amino acid sequence, its ability to attack and infect humans. If this adaption couldn't have happened in the community, the only place it could have happened is in a laboratory, by what is called serial passage, a common laboratory process that repeatedly gives the virus a chance to practice on humanized mice or VERO monkey cells.[10] A related study showing human adaption right from the start of the pandemic looked at which of the dozens of protein manufacturing tools that CoV-2 uses (called tRNAs). It showed the same uncanny adaptation to the human tools with no evidence that the tools from other potential intermediate hosts would be suitable.

This evidence presented makes a strong case that CoV-2 did not come from nature. But is there affirmative evidence that it could have come from a laboratory? The answer is yes.

Laboratory Origin Hypothesis

The spike protein that gives the coronavirus its name, corona or crown, is the key to match with the lock found in host cells. But before it can inject its genetic material in the host cell, the spike protein needs to be cut, to loosen it in preparation for infection. The host cell has the scissors or enzymes that do the cutting. The singular, unique feature of CoV-2 is that it requires a host enzyme called furin to activate it at a spot called the S1/S2 junction. No other coronavirus in the same subgenera has a furin cleavage site, as it is called. The other coronaviruses are cleaved at a site downstream from the S1/S2 site, called the S' site.

This is of course a major problem for the zoonosis theory, but it gets worse.

Since 1992 the virology community has known that the one sure way to make a virus deadlier is to give it a furin cleavage site at the S1/S2 junction in the laboratory. At least eleven gain-of-function experiments, adding a furin site to make a virus more infective, are published in the open literature, including Dr. Zhengli Shi, head of coronavirus research at the WIV. This has

[10] It is noteworthy that the furin cleavage site is actually unstable in passage in VERO cells and is often deleted within a few passages. A laboratory origin theory needs to account for this observation. On the other hand, mutations in the furin site among the human CoV-2 genomes are exceedingly rare.

caused a flurry of Chinese papers since the pandemic began trying to show a natural furin site in a related virus (this one example was later shown to be an error in interpretation) or to show that furin sites from distant cousins of CoV-2 might be the source through a process called recombination, where two different viruses infect the same host and then make a mistake in copying their genetic material, and swap sequences.

These convoluted, hypothetical methods each fail, however. It turns out that it is Daszak himself who has shown that the subgenera of coronaviruses that have furin sites are found in different bat hosts, which live in different regions of China, than the sarbecovirus subgenera of which CoV-2 is a member. And even with these barriers, they apparently are too far apart to recombine. "For the three focal subgenera, *Sarbecoviruses, Merbecoviruses and Embevoviruses*…none of the three focal subgenera recombines with one another."[11] As noted previously[2] Dr. Shi also does not believe the bats of Hubei province are capable of being a host for CoV-2-related coronaviruses.

But it gets worse still for the zoonosis theory. The gene sequence for the amino acids in the furin site in CoV-2 uses a very rare set of two codons, three letter words so six letters in a row, that are rarely used individually and have never been seen together in tandem in any coronaviruses in nature. But these same 'rare in nature' codons turn out to be the very ones that are always used by scientists in the laboratory when researchers want to add the amino acid arginine, the ones that are found in the furin site. When scientists add a dimer of arginine codons to a coronavirus, they invariably use the word, CGG-CGG, but coronaviruses in nature rarely (<1%) use this codon pair. For example, in the 580,000 codons of 58 Sarbecoviruses the only CGG pair is CoV-2; none of the other 57 sarbecoviruses have such a pair.[12]

So, there is no natural example of a furin protein site in nature that could be introduced into CoV-2 by recombination, there is no natural example of the particular gene sequence for the furin protein site contained in CoV-2 being used to code for anything in nature, but this particular coding is exactly what Dr. Shi, Baric, and others have used previously in published experiments to insert or optimize arginine codons.

It is telling that when Dr. Shi introduced the world to CoV-2 for the first time in January 2020 she showed hundreds of gene sequences of this novel virus but stopped just short of showing the furin site, the one she is purported to have introduced, seemingly not wanting to call attention to her handywork. She apparently failed to realize that an accomplished <u>but innocent</u> virologist, finding the first furin site ever seen in this class of viruses apparently coming from nature, would have featured the presence of the furin site prominently, and also would have used its presence and her experience with furin sites in other viruses to predict what it would foretell for the world due to its aggressive nature.

She could have perhaps saved many lives just by telling the world that she saw a furin site in the virus sequence. It would be left to a French and Canadian team to later identify the furin site in a

[11] CoV-2 is in the subgenera Sarbecoviruses. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009272
[12] https://virological.org/t/alignment-of-58-sarbecovirus-genomes-for-conservation-analysis-of-sars-cov-2/430

paper.[13] They would write: "This furin-like cleavage site…may provide a **gain-of-function** to the 2019-nCoV for efficient spreading in the human population compared to other lineage b betacoronaviruses." [Emphasis added.]

Dr. Shi has denied the virus came from her lab, but she has created such a record of multiple examples of obfuscation, half-truths, contrived specimens, genetic sequences taken from thin air but published in premier journals and US NIH databases, etc. that her veracity is deeply damaged. Perhaps her words and actions on December 30, 2019 show the truth. Her very first response when told there was an unknown outbreak in Wuhan and to return back quickly from a meeting she was attending in Shanghai was to say, "Could this have come from our lab?"[14]

"I wondered if [the municipal health authority] got it wrong," she says. "I had never expected this kind of thing to happen in Wuhan, in central China." Her studies had shown that the southern, subtropical provinces of Guangdong, Guangxi and Yunnan have the greatest risk of coronaviruses jumping to humans from animals—particularly bats, a known reservoir. After all, the US equivalent of the distance, climate change, and human population density change between Yunnan and Wuhan is comparing the Everglades National Park in Florida and New York City.

Her other action on December 30 was to alter WIV computer databases of novel coronaviruses used by the world's virologists for research to make it more difficult to search for which coronaviruses she had in her building. In short, the day she was asked to address the pandemic in Wuhan, she chose to spend time to make unavailable to her fellow scientists of the world her decades of coronavirus work.

The notion that CoV-2 was a laboratory creation, designed for maximum virulence, that escaped the laboratory accidentally has additional rings of evidence. From President Xi announcing in February new laws about laboratory security, to abundant evidence that the WIV was closed in October with few personnel inside, to the top military medical research doctor, General Chen Wei, being placed in charge of the WIV, to many more clues, it is clear an event occurred in Wuhan sometime in late 2019 that is most consistent with a laboratory escape.

The Asian region has a two-decade record of a little less than one laboratory-acquired infection per year. After the first SARS-CoV-1 epidemic was ended, SARS-CoV-1 jumped four more times into the human population, all from laboratories, with two in China. The last smallpox death in the entire world was a secretary who worked two floors above a research lab in England and contracted it through the ventilation system. The head of that laboratory committed suicide over his anguish for causing her death.

Over and over again. there is a long history and record of laboratory acquired infections that provides the background for considering what happened here.

[13] https://www.sciencedirect.com/science/article/pii/S0166354220300528?via%3Dihub

[14] https://www.scientificamerican.com/index.cfm/_api/render/file/?method=inline&amp;fileID=E1FDF8DE-9E22-4CE5-AD8B2E4682F52A86

Lab-made Bio-Weapon Hypothesis

But was SARS-CoV-2 more than just a gain-of-function experiment that escaped a laboratory? Could it have been one part of a two-part novel virus-vaccine bioweapons program?

General Chen Wei has been involved in vaccine research since joining the People's Liberation Army after college. In a 2017 internal speech at the AMMS (Academy of Military Medical Sciences) she said: *"只要有矛. 才能研究盾."* which translates roughly as, "you need to have an arrow to study a shield." I believe a Rubicon has been crossed by the world with this pandemic and framing the proper understanding of how we got here, and the proper response will be the critical next steps.

Evidence of adenovirus vaccine sequences in early patients would suggest both that SARS-CoV-2 was created in a laboratory and that there was sufficient priority set on this project to create a specific vaccine for the chimera coronavirus.

When Oppenheimer saw the application of Einstein's physics in the embodiment of the atomic bomb, he is said to have quoted a line from the Hindu scripture, the Bhagavad Gita, which reads: 'Now I am become Death, the destroyer of worlds.' The contribution of physics' research to human killing would total less than 300,000 people in two ten-square mile zones in Japan, and the horrors of those events led the world to regulate the raw materials of such bombs and to sanction sovereign nations who attempted to violate the rules.

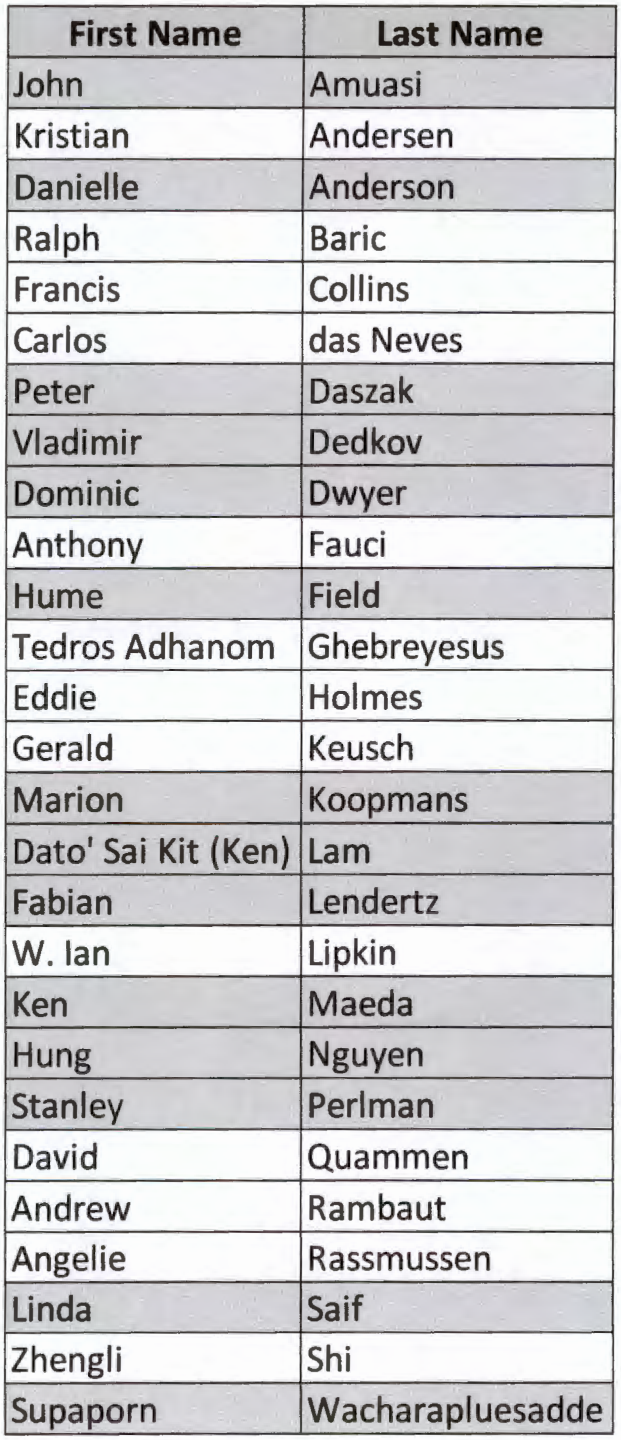
This had followed the contribution of chemistry to human killing in the form of chemical warfare during World War I, in which 100,000 were killed, and led the nations of the world to an historic agreement to never use chemical warfare again. It is now only 'rogue' operators who violate the norms civilized nations have agreed to.

It seems to be biology's turn to show its dark arts. If it is generally understood that biology/biotechnology has been harnessed to create a pandemic that has killed more people than physics and chemistry research combined, and to be a weapon where no place on earth is safe from its effects (SARS-CoV-2 has been detected in the deepest Amazon jungles and at research stations in Antarctica), there needs to be developed a new set of regulations, rules, etc. to both honor the 1.8 million innocent people who died from COVID-19 and to protect the world so this never happens again. It is also urgent to gather further data to support or refute if this was a Chinese bioweapons program, as the consequences of that would be significant.

Pre-publication peer review. The manuscript was provided by email to the following medical and scientific peers to afford an opportunity to review, comment, and critique the manuscript before publication. Those highlighted in yellow are members of the WHO-convened Global Study of the Origins of SARS-CoV-2[15], The Lancet COVID-19 Commission[16], or both.

[15] https://www.who.int/health-topics/coronavirus/origins-of-the-virus

[16] https://covid19commission.org/origins-of-the-pandemic



First Name	Last Name
John	Amuasi
Kristian	Andersen
Danielle	Anderson
Ralph	Baric
Francis	Collins
Carlos	das Neves
Peter	Daszak
Vladimir	Dedkov
Dominic	Dwyer
Anthony	Fauci
Hume	Field
Tedros Adhanom	Ghebreyesus
Eddie	Holmes
Gerald	Keusch
Marion	Koopmans
Dato' Sai Kit (Ken)	Lam
Fabian	Lendertz
W. Ian	Lipkin
Ken	Maeda
Hung	Nguyen
Stanley	Perlman
David	Quammen
Andrew	Rambaut
Angelie	Rassmussen
Linda	Saif
Zhengli	Shi
Supaporn	Wacharapluesadde

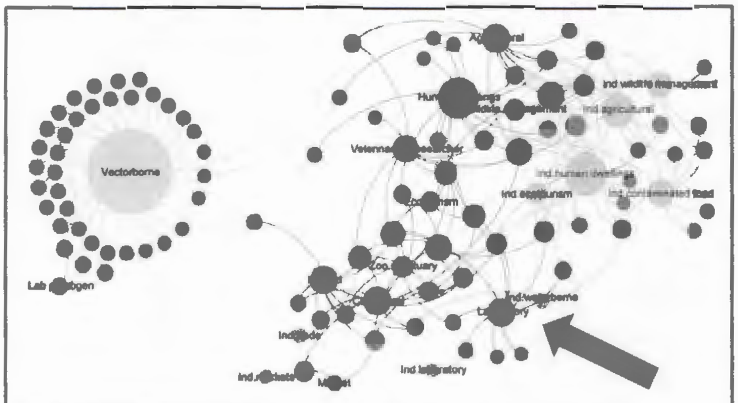
A Bayesian analysis concludes beyond a reasonable doubt that SARS-CoV-2 is not a natural zoonosis but instead is laboratory derived

Introduction. A two-hypothesis, Bayesian analysis was conducted to determine the origin of the SARS-CoV-2 pandemic. The conclusion was that it was created in a laboratory with synthetic biology tools from a bat beta coronavirus, subgenera sarbecovirus backbone (98.9% probability) and not from a natural, zoonotic transmission (1.1%).

There is no direct evidence of whether the release was accidental, or deliberate but circumstantial evidence makes it is highly likely it was accidental.

At the one-year anniversary of the first cases of COVID-19, the coronavirus pandemic caused by the SARS-CoV-2 virus, the origin of the virus remains unknown. While leading institutions and experts have been consistently adamant that it is a zoonotic disease which jumped from a bat reservoir host to humans directly or through an intermediate host the alternative possibility that it escaped from a laboratory conducting research remains a viable option.

In fact, in 2015 Peter Daszak, a leading zoonotic proponent of CoV-2 origin, wrote in, "Spillover and pandemic properties of zoonotic viruses with high host plasticity,"[17] that transmission from laboratories was a major source of zoonotic disease. The Figure below from the Daszak paper shows this important relationship (green arrow):



Epidemiologic bipartite network map showing high-risk disease transmission interfaces shared by zoonotic viruses transmitted from wildlife to humans.

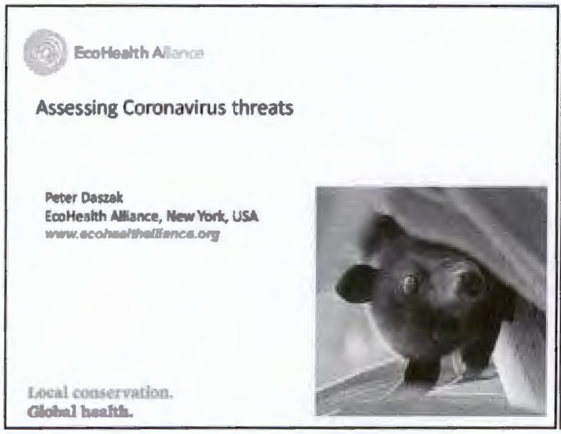
High-risk interfaces are shown with node size proportionate to the number of viruses reported for each transmission interface, categorized according to (1) direct contact with wildlife (dark blue), (2) indirect contact with wildlife (light blue) and (3) transmission by vector (yellow). Virus node size (red, n = 86 viruses) reflects the number of connections to different transmission interfaces and ecological plasticity of viruses through use of multiple transmission opportunities. Highly connected and more central interfaces facilitated transmission of more viruses, providing an epidemiologic picture of circumstances likely to promote future disease emergence and important targets for disease surveillance and preventive measures.

Daszak et al. also writes: "**Zoonotic virus spillover** from wildlife was most frequent in and around human dwellings and in agricultural fields, as well as **at interfaces with occupational exposure to animals** (hunters, **laboratory workers**, veterinarians, **researchers**, wildlife management, zoo and sanctuary staff). **Primate hosts were most frequently cited as the source of viruses transmitted by direct contact** during hunting (exact P = 0.051) and **in laboratories**

[17] https://www.nature.com/articles/srep14830

(exact P = 0.009)." [Emphasis added]. Primate "hosts" can presumably include monkey cell culture, such as the ubiquitous VERO cell used in all virology laboratories, including the WIV.

In 2015 Dr. Daszak spoke of the spillover danger of certain types of laboratory research:

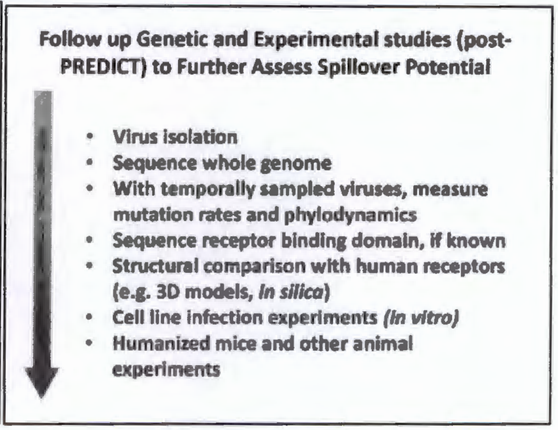


EcoHealth Alliance

Assessing Coronavirus threats

Peter Daszak
EcoHealth Alliance, New York, USA
www.ecohealthalliance.org

Local conservation.
Global health.

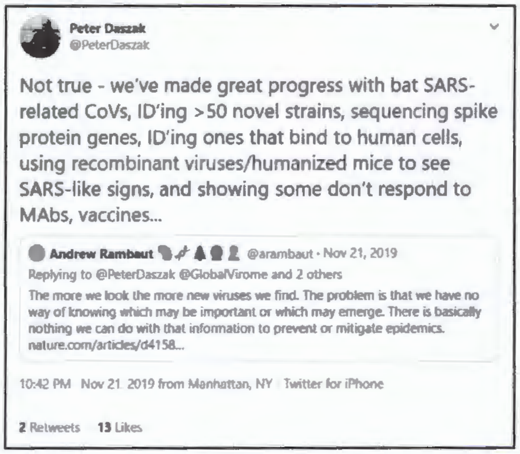


Follow up Genetic and Experimental studies (post-PREDICT) to Further Assess Spillover Potential

- Virus isolation
- Sequence whole genome
- With temporally sampled viruses, measure mutation rates and phylodynamics
- Sequence receptor binding domain, if known
- Structural comparison with human receptors (e.g. 3D models, *in silico*)
- Cell line infection experiments *(in vitro)*
- Humanized mice and other animal experiments

He writes: "with each step, increased risk possible" with "Humanized mice and other animal experiments" the highest risk work.

In a prescient Twitter post in November 2019, he highlights the work he is doing using recombinant viruses with humanized mice and making viruses that **"don't respond to MAbs, vaccines…"** in response to criticism his work is of limited value:



Peter Daszak
@PeterDaszak

Not true - we've made great progress with bat SARS-related CoVs, ID'ing >50 novel strains, sequencing spike protein genes, ID'ing ones that bind to human cells, using recombinant viruses/humanized mice to see SARS-like signs, and showing some don't respond to MAbs, vaccines...

Andrew Rambaut @arambaut · Nov 21, 2019
Replying to @PeterDaszak @GlobalVirome and 2 others
The more we look the more new viruses we find. The problem is that we have no way of knowing which may be important or which may emerge. There is basically nothing we can do with that information to prevent or mitigate epidemics. nature.com/articles/d4158...

10:42 PM · Nov 21, 2019 from Manhattan, NY · Twitter for iPhone

2 Retweets 13 Likes

Clearly, before the beginning of the pandemic, Daszak, now a member of both the WHO and Lancet teams being sent to China to explore the origin of CoV-2, could entertain the eal possibility of a laboratory created virus escaping into the human population/community.

The purpose of this analysis is to use a Bayesian Inference Network approach to the collected circumstantial evidence that is available to provide likelihoods of the alternative hypotheses as to the origin of SARS-CoV-2. The analysis also will include certain prior probabilistic conclusions to help set the initial state before the proprietary evidence is used.

Origin hypotheses: Initial States to establish the posterior probabilities.

Two published Bayesian analyses and two independent studies of zoonotic spillover from nature and laboratory-acquired infections in Asia will be used to establish the posterior probabilities for this analysis.

Zoonotic spillover frequency versus laboratory acquired infection frequency based on two published papers, one by Daszak et al.

In 2015 Daszak et al. published a paper entitled, "Spillover and pandemic properties of zoonotic viruses with high host plasticity,"[1] in which they identified 162 zoonotic viruses with naturally occurring animal-to-human transmission from 1990-2010. This is a frequency of 162/20 = 8.1 events per year.

They also note: "The majority (94%) of zoonotic viruses described to date (n = 162) are RNA viruses, which is 28 times higher (95% CI 13.9–62.5, exact P < 0.001) than the proportion of RNA viruses among all vertebrate viruses recognized, indicating that RNA viruses are far more likely to be zoonotic than DNA viruses." CoV-2 is an RNA virus.

Finally, they note that: "In general, wild animals were suggested as the source of zoonotic transmission for 91% (86/95) of zoonotic viruses compared to 34% (32/95) of viruses transmitted from domestic animals and 25% (24/95) with transmission described from both wild and domestic animals."

One of the caveats of the Daszak data is that it categorizes a laboratory-acquired infection (LAI) from an animal collected from the wild as a zoonotic spillover. There is no data in the paper to assess this issue and leaving it uncorrected is a conservative approach since it only inflates the natural zoonotic frequency.

In 2018 a paper by Siengsanan-Lamont entitled, "A Review of Laboratory-Acquired Infections in the Asia-Pacific: Understanding Risk and the Need for Improved Biosafety for Veterinary and Zoonotic Diseases," was published.[18] They reported 27 LAIs between 1982 and 2016, a frequency of 27/(2016 – 1982) = 0.8 events per year.

Using these historical frequencies of zoonotic spillover versus LAI to predict a future event can be calculated in the following manner:

Evidence	Zoonotic Origin	Laboratory Origin
Frequency per year from Daszak paper	8.1	NA
Frequency per year from Siengsanan-Lamont paper	NA	0.8
Total events per year	8.1 + 0.8 = 8.9	8.1 + 0.8 = 8.9
Likelihood of future event based on historical frequency	8.1/8.9 X 100 = 0.91	0.8/8.9 X 100 = 0.9

Daszak's initial state analysis. This evidence sets the likelihood that CoV-2 was a zoonotic origin event at 91% and a laboratory origin event at 9%.

[18] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6073996/

Independent prior analyses: Rootclaim.

The next data that will be used is a recent analysis published on the Rootclaim website.[19] Three hypotheses below were analyzed through a series of evidence statements and the probabilities that each was the origin of SARS-CoV-2 determined:

Hypothesis	Calculated Probability
Lab escape: The virus was the subject of genetic research, including gain-of-function, and was released by accident	81%
Zoonotic: The virus evolved in nature and was transmitted to humans from a non-human vertebrate animal	16%
Bioweapon: The virus was genetically engineered as a bioweapon and was deliberately released	3%

As can be seen, the highest likelihood probability is an accidental lab escape, the lowest a bioweapon. The details of the evidence used to arrive at this conclusion is contained in Appendix 1. A summary of the changes in probability at each level of evidence analysis is shown in this table:

Evidence	Laboratory	Zoonosis	Bioweapon
Starting point	1.2%	82%	16%
Contagion and mortality	1.4%	97%	1.9%
Outbreak location: Wuhan	42%	56%	2.8%
Virus sources near Wuhan	16%	83%	1.0%
Chimera	37%	60%	2.5%
Furin cleavage	72%	23%	4.8%
WIV lab procedures	80%	17%	3.5%
WIV disassociation	89%	9%	2.0%
Chinese response	90%	8%	1.7%
No reported infections at WIV	86%	11%	2.4%
No whistleblowers	81%	16%	2.8%

As can be seen, the starting point assumed an 82% probability of a zoonotic origin. This starting point is a reasonable value and will be used here. Since some of the evidence in the above analysis will be used here, only the starting point will be used and not the probability changes from there.

For purposes of this analysis only the Rootclaim initial state will be used since much of their evidence is also covered in the analysis here.

[19] https://www.rootclaim.com/analysis/what-is-the-source-of-covid-19-sars-cov-2

In a paper by Daszak and colleagues it states: "In general, wild animals were suggested as the source of zoonotic transmission for 91% (86/95) of zoonotic viruses compared to 34% (32/95) of viruses transmitted from domestic animals and 25% (24/95) with transmission described from both wild and domestic animals."[1]

On the other hand, domestic animals seem to have been ruled out for SARS-CoV-2. In an interview for *Science* in July 2020, Dr. Zhengli Shi, head of coronavirus research at the Wuhan Institute of Virology, stated: "Under the deployment of the Hubei Provincial Government, our team and researchers from Huazhong Agricultural University collected samples of farmed animals and livestock from farms around Wuhan and in other places in Hubei Province. We did not detect any SARS-CoV-2 nucleic acids in these samples."[20]

Reanalysis of Rootclaim initial state to remove Bioweapons option.

The US government uses the following definitions:

"Gain-of-function (GOF) studies, or research that improves the ability of a pathogen to cause disease, help define the fundamental nature of human-pathogen interactions, thereby enabling assessment of the pandemic potential of emerging infectious agents, informing public health and preparedness efforts, and furthering medical countermeasure development.

Gain-of-function studies may entail biosafety and biosecurity risks; therefore, the risks and benefits of gain-of function research must be evaluated, both in the context of recent U.S. biosafety incidents and to keep pace with new technological developments, in order to determine which types of studies should go forward and under what conditions."[21]

"Dual use research of concern (DURC) is life sciences research that, based on current understanding, can be reasonably anticipated to provide knowledge, information, products, or technologies that could be directly misapplied to pose a significant threat with broad potential consequences to public health and safety, agricultural crops, and other plants, animals, the environment, materiel, or national security. "[22]

For this analysis, the assumption is made that GOF and DURC are largely the same processes and techniques in the laboratory and thus can only be distinguished by direct, documentary evidence of the intent of the research from administers in the facilities conducting the work.

In the absence of any such documentary evidence that bioweapon research was being conducted or that SARS-CoV-2 is a bioweapon and to take the least inflammatory posture, the initial state for the above prior analysis will be recalculated by eliminating the hypothesis, and its accompanying probability, that SARS-CoV-2 was created as a bioweapon. The revised initial state calculation is shown in this table:[23]

[20] https://www.sciencemag.org/sites/default/files/Shi%20Zhengli%20Q%26A.pdf

[21] https://www.phe.gov/s3/dualuse/Pages/GainOfFunction.aspx

[22] https://www.phe.gov/s3/dualuse/Pages/default.aspx

[23] For clarity, the 3% bioweapon probability was simply dropped and the remaining likelihoods, 81% and 16%, were normalized.

Evidence	Zoonotic Origin	Laboratory Origin	Bioweapons Origin
Rootclaim initial state	0.86	0.012	0.16
Remove bioweapons	NA	NA	0
Normalize remaining hypotheses	0.86/(0.86 + 0.012) = 0.986	0.012/(0.86 + 0.012) = 0.014	NA

Rootclaim Initial state analysis, adjusted. This evidence sets the likelihood that CoV-2 was a zoonotic origin event at 98.6% and a laboratory origin event at 1.4%.

Additional Prior Evidence by Demaneuf and De Maistre. A second prior Bayesian analysis was performed by professionally educated risk assessment personnel and Chinese-language speaking professionals[24] and is included herein in its entirety. For the sake of brevity, the zoonotic origin evidence was based primarily on population size, distribution, and geographic distribution of bat populations relative to Wuhan. With respect to a lab accident, they separately analyze probabilities of a virus escape during collection, transport, and direct lab accidents and then separately the probability of a community outbreak following a lab escape. They also use primary Mandarin-language sources for Chinese estimates of the same events, showing corroboration of the probabilities. Their conclusion is that the probability of a lab escape ranges from 6% to 55% with a zoonotic origin a zoonotic origin probability being 45% to 94%.

Second Bayesian analysis. Using the most conservative probabilities, this evidence sets the likelihood that CoV-2 was a zoonotic origin event at 94% and a laboratory origin event at 6%.

Selection of initial state for Bayesian analysis.

The Text-Table below summarizes the three approaches to an initial state as to the origin of CoV-2. While the Demaneuf and De Maistre analyses set a range for the zoonotic origin of 45% to 94%, I have used the top of the range of their probability of a zoonotic origin to be conservative.

Prior Analysis	Zoonotic Origin	Laboratory Origin
Daszak et al. paper	91%	9%
Rootclaim Bayesian analysis	98.6%	1.4%
Demaneuf and De Maistre Bayesian analysis	94%	6%

Using a simple online calculator[25] the mean of these three value sets is 94.5%, the standard deviation is ± 3.8%, and the 95% confidence interval is ± 4.3%. Using these data, the upper bound of the 95% confidence interval is 98.8% and, to be most conservative, this will be used as the starting probability of a zoonotic origin.

Initial state for this analysis. The likelihood that SARS-CoV-2 began as a zoonotic event is 98.8% and the likelihood it began as a laboratory event is 1.2%.

[24] https://zenodo.org/record/4067919#.X-qlm9gzbOj . For reference purposes, this paper comes with a spreadsheet listing 112 individual BSL-3 labs in China across 62 lab-complexes.

[25] https://www.calculator.net/standard-deviation-calculator.html?numberinputs=91%2C+94%2C+98.6&ctype=s&x=48&y=19

1. General approach of this analysis[26]

This analysis is intended to examine two competing and mutually exclusive theories of the origin of the coronavirus, SARS-CoV-2 (CoV-2), and the pandemic it has caused, COVID-19.

At the time of this writing there have been 83 million confirmed cases and 1.8 million deaths.[27] Some sources place the economic damage at $21 trillion USD.

Bayes Theorem

This brief description of the Bayes Theorem was taken from the work of Jon Seymour:[28]

"The eponymously named Bayes Theorem was discovered by the Reverend Thomas Bayes in the 1700's and saved for posteriority by an archivist of his papers who discovered the work posthumously. In common language, it provides a rational technique for revising a prior belief in light of new evidence. The equation for Bayes Theorem is given below:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

where:

- H is the statement of the hypothesis of interest
- P(H) is the prior probability that the hypothesis is true, independent of the evidence.
- E is the evidence being used to revise the belief in hypothesis
- P(E) is the marginal likelihood of the evidence, independent of the hypothesis
- P(E|H) is the likelihood the evidence, given that the hypothesis is true
- P(H|E) is the posterior probability of the hypothesis, given the evidence.

P(E) is sometimes difficult to estimate, but the following identity must hold:

$$P(E) = P(E|H).P(H) + P(E|\widehat{H}).P(\widehat{H})$$

Here P(E|^H) is the probability of the evidence, assuming the hypothesis is false and P(^H) is the probability the hypothesis is false which is the same as 1-P(H). Estimating the two conditional probabilities P(E|H) and P(E|^H) is generally easier than estimating the unconditional probability, P(E)."

[26] The statistical approach and many of the individual statistical analyses were performed by Dr. Martin Lee, PhD, Adjunct Professor of Biostatistics, UCLA. https://ph.ucla.edu/faculty/lee The likelihood adjustments to the Bayesian analysis, which you can see are routine math, were conducted by the author.

[27] https://www.worldometers.info/coronavirus/coronavirus-cases/

[28] https://jonseymour.medium.com/a-bayesian-analysis-of-one-aspect-of-the-sars-cov-2-origin-story-where-the-first-recorded-1fbdcbea0a2b

Theory One. The zoonotic theory is that a vertebrate animal was infected with CoV-2 or an ancestor (Index Host) and that a human was infected with contact to that Index Host in some manner. Human-to-human spread then followed.

Theory Two. The laboratory origin theory is that CoV-2 or an ancestor was being used in laboratory experiments and that it 'escaped' from the lab via an infected person, lab animal, experimental waste, etc.

I have found no evidence of a deliberate release and early firsthand accounts of local officials and scientists suggest surprise and consternation. If this was a deliberate release, such evidence would be extremely local, limited in distribution, and highly compartmentalized. It is beyond the scope of this analysis.

Weight of the evidence. For purposes of the calculation of posterior probabilities in the Bayesian analysis, evidence which has a statistical basis will be used directly to adjust the probabilities.

Statistically significant evidence. Since some of the probability calculations have astronomical values which would make a single such evidence statement, if inputted directly, swamp any further calculation and make their later contribution mute, a decision was made to simply treat quantitative probabilities as significant at the p = 0.05 level, no matter how much 'more significant' the calculation suggested.

So, for example, a probability of certain codon usage coming from nature may be one in 440 or p = 0.002, the contribution of this evidence to the input to the posterior probability adjustment would be set at a p-value of 0.05. In such cases the adjustment would be to change the 'winning' hypothesis by multiplying by 19, since a p = 0.05 is the same as a 19 out of 20 likelihood event. This is a conservative treatment of what would be highly significant data.

Other quantitative evidence. If a piece of evidence can be quantified but it does not reach a significance of p = 0.05 it will be used directly in the likelihood adjustment.

Non-quantitative evidence. For evidence that cannot be quantified, the decision was made to treat these as quantitative outcomes with a 51% to 49% likelihood value with respect to the 'winning' hypothesis. This has the effect of increasing the probability of that hypothesis for that step in the Bayesian analysis by 1.04. This 51%/49% concept is related to the legal standard of the 'preponderance of the evidence' used in civil litigation.

Independence. An important qualitative assessment that must be made is whether or not two pieces of evidence are independent of each other. If they are independent, they can each be used in determining a new likelihood calculation. If they are dependent on each other then they must be combined and only a single new likelihood analysis can be made. Where ever possible, evidence statements that could be considered as dependent are called out and this rule is followed on their contribution to the analysis.

Subjective Discount Factor. The impact of each piece of evidence was adjusted further by a subjective discount factor. This is a qualitative assessment of the overall veracity of a particular piece of evidence when all factors, samples, methods, data sources, etc. are taken into context. It varies from 60% to 100% and is used as a fraction to reduce the impact of a single piece of evidence even further.

Hearsay. Just as in a court of law, evidence, usually attributed to a given person or persons, that is not directly available but instead relies on statements of others is usually not allowed in a court trial and will accordingly not be used here to adjust the Bayesian analysis. It may be recorded and preserved as a placeholder and reminder for further research. If new, direct evidence can be found than the bar of using it is lifted and it can be used for adjustment.

Significant figures. Because of the overall nature of the analyses here, all math calculations related to likelihoods are performed and carried forward at the 'one significant figure' level, with standard rounding rules applied. This has the effect, near the end of the cumulative evidence, of failing to change the relative probabilities as the small adjustments are reversed in the rounding process.

Evidence. International committees to investigate the origin of SARS-CoV-2 may not be impartial.

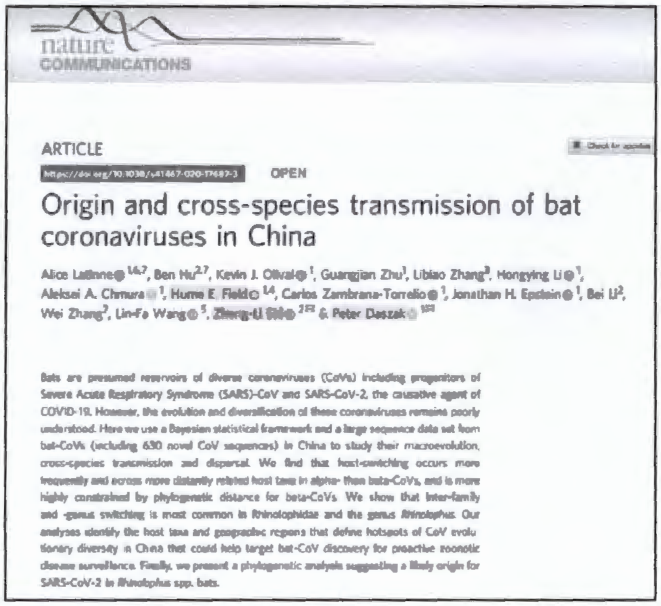
At the time of the writing of this manuscript there are two committees charged with examining the evidence and determining the origin of the SARS-CoV-2 virus. One committee is commissioned by the World Health Organization (WHO) and the other is an ad hoc committee established by the British medical journal, *The Lancet.*

The composition of the two committees is shown in the Text-Table below:

	Lancet Commission of CoV-2	**WHO Commission on CoV-2 origin**
	Dr. Peter Daszak, Chair	**Dr. Peter Daszak, Ph.D (EcoHealth Alliance, USA)**
	Dr. John Amuasi	Prof. John Watson (Public Health England, United Kingdom)
	Dr. Danielle Anderson	Prof. Dr. Marion Koopmans, DVM PhD (Erasmus MC, Netherlands)
	Dr. Isabella Eckerle	Prof. Dr. Dominic Dwyer, MD (Westmead Hospital, Australia)
Also co-author	Dr. Hume Field	Vladimir Dedkov, Ph.D (Institute Pasteur, Russia)
	Dr. Gerald Keusch	Dr. Hung Nguyen, PhD (International Livestock Research Institute (ILRI), Vietnam)
	Dr. Dato' Sai Kit (Ken) Lam	PD. Dr. med vet. Fabian Lendertz (Robert Koch-Institute, Germany)
	Dr. Carlos das Neves	Prof. Dr. Thea Fisher, MD, DMSc(PhD) (Nordsjællands Hospital, Denmark)
	Dr. Malik Peiris	Dr. Farag El Moubasher, Ph.D (Ministry of Public Health, Qatar)
	Dr. Stanley Perlman	Prof. Dr. Ken Maeda, PhD, DVM (National Institute of Infectious Diseases, Japan)
	Dr. Linda J. Saif	WHO Commission of CoV-2 origin
	Dr. Supaporn Wacharapluesadee	
	Lancet Commission on CoV-2	
	Signed Lancet letter	
	Co-author with Daszak	

There are a number of potential conflicts of interest:

Fully half of The Lancet's team had already suggested that any lab-leak hypothesis was a "conspiracy theory" in a January 2020 paper that has been shown elsewhere within to have been orchestrated behind the scenes to appear spontaneous.



nature
COMMUNICATIONS
Check for updates

ARTICLE

https://doi.org/10.1038/s41467-020-17687-3 OPEN

Origin and cross-species transmission of bat coronaviruses in China

Alice Latinne[1,6,7], Ben Hu[2,7], Kevin J. Olival[1], Guangjian Zhu[1], Libiao Zhang[3], Hongying Li[1], Aleksei A. Chmura[1], Hume E. Field[1,4], Carlos Zambrana-Torrelio[1], Jonathan H. Epstein[1], Bei Li[2], Wei Zhang[2], Lin-Fa Wang[5], Zheng-Li Shi[2✉] & Peter Daszak[1✉]

Bats are presumed reservoirs of diverse coronaviruses (CoVs) including progenitors of Severe Acute Respiratory Syndrome (SARS)-CoV and SARS-CoV-2, the causative agent of COVID-19. However, the evolution and diversification of these coronaviruses remains poorly understood. Here we use a Bayesian statistical framework and a large sequence data set from bat-CoVs (including 630 novel CoV sequences) in China to study their macroevolution, cross-species transmission and dispersal. We find that host-switching occurs more frequently and across more distantly related host taxa in alpha- than beta-CoVs, and is more highly constrained by phylogenetic distance for beta-CoVs. We show that inter-family and -genus switching is most common in Rhinolophidae and the genus *Rhinolophus*. Our analyses identify the host taxa and geographic regions that define hotspots of CoV evolutionary diversity in China that could help target bat-CoV discovery for proactive zoonotic disease surveillance. Finally, we present a phylogenetic analysis suggesting a likely origin for SARS-CoV-2 in *Rhinolophus* spp. bats.

The above paper published in August 2020 has as co-authors Drs. Hume, Daszak, and Shi. Having two of these scientists be asked to investigate a third co-author is a clear conflict of interest.

A newspaper piece about Peter Daszak entitled, "The doctor who denied COVID-19 was leaked from a lab had this major bias,"[29] questions his ability to be unbiased due to a deep, long history of work with Dr. Zhengli Shi of the WIV.

A lengthy piece in Wired was subtitled, "The two major investigations into the origins of the pandemic are compromised by potential conflicts of interest."[30]

Since the purpose of this manuscript is to evaluate the scientific evidence concerning the origin of SARS-CoV-2 no further effort will be put into these matters. If and when a report is prepared from either committee there will be time to analysis the work in the reports and compare it to prior publications and statements from the committee members to look for bias.

Likelihood from initial state is unchanged following this evidence analysis:

Zoonotic origin (98.8%) and laboratory origin (1.2%)

[29] https://nypost.com/2021/01/16/doctor-who-denied-covid-was-leaked-from-a-lab-had-this-major-bias/

[30] https://www.wired.com/story/if-covid-19-did-start-with-a-lab-leak-would-we-ever-know/?utm_source=twitter&utm_medium=social&utm_campaign=onsite-share&utm_brand=wired&utm_social-type=earned

Evidence. Three high visibility papers grounded the zoonotic origin hypothesis in the public conversation from February to May 2020: a pros and cons analysis.

Introduction. The two key data points from December 2019 concerning the origin of the SARS-CoV-2 coronavirus infection, the cause of COVID-19, are the observation that a large number of the earliest patients worked or had visited the Hunan Seafood Market in Wuhan, China and that the hospitals where the first patients were admitted were a short distance from the Wuhan Institute of Virology (WIV), the only high security, BSL-4 laboratory in all of China, and arguably the leading research institute in the world studying coronaviruses of the type causing COVID-19.

The first data point is reminiscent of the origin of SARS-CoV-1, a zoonosis with interspecies transmission from bats to civet cats and then to humans, identified in wet markets in southern China. The second data point is reminiscent of the four SARS-CoV-1 human spillovers that occurred after the 2003 epidemic ended and were each a laboratory-acquired infection (LAI) by a scientist working in a government research laboratory, much like the WIV, and then local human-to-human spread and nearby hospital admission.

To be clear in this paper, the term zoonosis will only be used to describe a interspecies transmission outside of a laboratory. This point seems important to clarify since Dr. Zhengli Shi, head of coronavirus research at the WIV, has previously reported: "An outbreak of hemorrhagic fever with renal syndrome occurred among students in a college (College A) in Kunming, Yunnan province, China in 2003. Subsequent investigations revealed the presence of hantavirus antibodies and antigens in laboratory rats at College A and two other institutions. Hantavirus antibodies were detected in 15 additional individuals other than the index case in these three locations. Epidemiologic data indicated that the human infections were a result of **zoonotic transmission** of the virus from laboratory rats."[31] [emphasis added.] The author has found no other support for the use of the term zoonotic transmission with respect to an LAI and its dual use could be confusing, and so will be avoided.

While the two initial data points would suggest that a balanced approach should be taken with respect to investigations of the origin of SARS-CoV-2, three high visibility publications that argued the laboratory origin idea was a "conspiracy theory" and strongly argued that it was of zoonotic origin foreclosed legitimate debate for much of 2019. The purpose of this evidence analysis is to examine these papers and weigh the strength of the evidence.

Paper 1: The February 3, 2020 paper by WIV scientist Dr. Shi et al. entitled: "A pneumonia outbreak associated with a new coronavirus of probable bat origin."

This seminal paper set the stage for the zoonotic origin of SARS-CoV-2 and has been accessed over one million times. According to *Nature*, this article is in the 99th percentile (ranked 24th) of the 326,159 tracked articles of a similar age in all journals and the 99th percentile (ranked 2nd) of the 783 tracked articles of a similar age in Nature.

[31] https://pubmed.ncbi.nlm.nih.gov/20380897/

However, a careful analysis of it shows serious issues which suggest it is unreliable. The following analysis is in the form of an independent manuscript:

The seminal paper from the Wuhan Institute of Virology claiming SARS-CoV-2 probably originated in bats appears to contain a contrived specimen, an incomplete and inaccurate genomic assembly, and the signature of laboratory-derived synthetic biology

The coronavirus RaTG13 was purportedly identified in a bat "fecal" specimen that is probably not feces, has significant unresolved method-dependent genome sequence errors and an incomplete assembly with significant gaps, and has an anomalous base substitution pattern that has never been seen in nature but is routinely used in codon-optimized synthetic genome constructions performed in the laboratory

Abstract. The species of origin for the SARS-CoV-2 coronavirus that has caused the COVID-19 pandemic remains unknown after over six months of intense research by investigators around the world. The current consensus theory among the scientific community is that it originated in bats and transferred to humans either directly or through an intermediate species; no credible intermediate species exists at this time. The suggested origin early on from a Wuhan "wet market" has been determined to be a red herring and the pangolin is no longer considered a likely intermediate by the virology community.

The basis for the hypothesis that SARS-CoV-2 probably evolved from bats initially came from a February 2020 paper[32] from Dr. Zheng-Li Shi's laboratory at the Wuhan Institute of Virology (WIV). In that paper the Wuhan laboratory made two claims: 1), "a bat fecal sample collected from Tongguan town, Mojiang county in Yunnan province in 2013" contained a coronavirus, originally designated "Rhinolophus bat coronavirus BtCoV/4991[33]" in 2016 but renamed in their paper, RaTG13; and 2), the genomes of RaTG13 and SARS-CoV-2 had an overall identity of 96.2%, making it the closest match to SARS-CoV-2 of any coronavirus identified at that time. RaTG13 remains the closest match to SARS-CoV-2 at the current time.

In this paper I document that:

1) The RaTG13 specimen was not a bat fecal specimen, based on a comparison of the relative bacterial and eukaryotic genetic material in the purported fecal specimen to nine authentic bat fecal specimens collected in the same field visits as RaTG13 was collected by the Wuhan laboratory, run on the same Illumina instrument (id ST-J00123), and published in a second paper in February 2020.[15] While the authentic bat fecal

[32] Zhou, P., Yang, X., Wang, X. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579,** 270–273 (2020). https://doi.org/10.1038/s41586-020-2012-7 .

[33] A Coronavirus BtCoV/4991 Genbank entry by Dr. Shi records: organism="Rhinolophus bat coronavirus BtCoV/4991." In July 2020 she wrote: "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

samples were, as expected, largely bacterial (specifically, 65% bacteria and 12% eukaryotic genetic sequences), the purported RaTG13 specimen had a reversed composition, with mostly eukaryotic genes and almost no bacterial genetic material (0.7% bacteria and 68% eukaryotic). The RaTG13 specimen was also only 0.01% virus genes compared to an average of 1.4% for authentic bat fecal specimens. A Krona analysis identified 3% primate sequences consistent with VERO cell contamination, the standard monkey cell culture used for coronavirus research, including at the Wuhan laboratory. Based on using the mean and standard deviation of the nine authentic bat fecal specimens from the Wuhan laboratory, the probability that RaTG13 came from a true fecal sample but had the composition reported by the Wuhan laboratory is one in thirteen million;

2) According to multiple references, RaTG13 was identified via Sanger dideoxy sequencing before 2016, partially sequenced by amplicon sequencing in 2017 and 2018, and then complete sequencing and assembly by RNA-Seq in 2020, although some reports from WIV suggest the timing of the RNA-Seq experiments may have been performed earlier than 2020. In any case, a Blast analysis of sequences from the amplicon and RNA-Seq experiments indicates an approximate 5% nucleotide difference, 50-fold higher than the technical error rate for RNA-Seq of about 0.1%. At least two gaps of over 60 base-pairs, with no coverage in the RNA-Seq data, were easily identified. The incomplete assembly and anomalous, method-dependent sequence divergence for RaTG13 is troublesome;

3) The pattern of synonymous to non-synonymous (S/NS) sequence differences between RaTG13 and SARS-CoV-2 in a 2201 nucleotide region flanking the S1/S2 junction of the Spike Protein records 112 synonymous mutation differences with only three non-synonymous changes. Based on the S/NS mutational frequencies elsewhere in these two genomes and generally in other coronaviruses the probability that this mutation pattern arose naturally is approximately one in ten million. A similar pattern of unnatural S/SN substitutions was seen in a 10,818 nt region of the pp1ab gene. This pp1ab gene pattern has a probability of occurring naturally of less than one in 100 billion. A total of four regions of the RaTG13 genome, coding for 7,938 nt and about one-quarter of the entire genome, contain over 200 synonymous mutations without a single non-synonymous mutation. This has a probability of one in 10^{-17}. A possible explanation, the absolute criticality of the specific amino acid sequence in the regions which might make a non-synonymous change non-infective, is ruled out by the rapid appearance of an abundance of non-synonymous mutations in these very regions when examining the over 80,000 human SARS-CoV-2 specimens sequenced to date. An alternative hypothesis, that this arose by codon substitution is examined. It is demonstrated, by example from a published codon-optimized SARS-Cov-2 Spike Protein experiment, that the anomalous S/SN pattern is precisely the pattern which is produced, by design, when synthetic biology is used and represents a signature of laboratory construction.

Based on the findings concerning the RaTG13 data, including anomalies and inconsistent statements about RaTG13, its origin, renaming, and sequencing timing; the finding that the specimen it is purported to have come from is not bat feces and has a signature of cell culture contamination; the unexplained method-dependent 5% sequence difference for RaTG13; and the S/SN mutation pattern reported, which to my knowledge has never been seen in nature, it can be concluded that RaTG13 is not a pristine biological entity but shows evidence of genetic manipulation in the laboratory.

Until a satisfactory explanation of the findings in this paper have been offered by the Wuhan laboratory, all hypotheses of the proximal origin of the entry of SARS-CoV-2 into the human population should now include the likelihood that the seminal paper contains contrived data. For example, the hypothesis that SARS-CoV-2 was the subject of laboratory research and at some point escaped the laboratory should be included in the narrative of the origin of SARS-CoV-2 research.

Introduction. Since the first reported patient on December 1, 2019 with a SARS-CoV-2 infection, the virus has caused a pandemic that has led to twenty-five million cases worldwide and over 840,000 deaths as of August 30, 2020. To make progress on treating this disease and preventing the next viral outbreak, knowing the origin of the virus and how it entered the human population is critical.

On February 3, 2020 a paper was published from the Wuhan Institute of Virology that identified a bat coronavirus, RaTG13, as having a 96.2% identity to SARS-CoV-2, quickly providing support for a zoonotic origin, either from bats directly or from bats to humans through an unknown intermediary species. If true, this would replicate the model of SARS-CoV 2003 in which the transmission was from bats to civets to humans and for MERS in which the transmission was from bats to camels to humans. At the time of this paper and through August 30, 2020, no other virus has been identified with a closer sequence homology to SARS-CoV-2 than RaTG13. The publication containing the RaTG13 sequence has been cited over 1600 times in the six months since publication. None of these studies contain research on the isolated virus itself since the virus has never been isolated or cultured. It was apparently found in only one sample from 2013 and that sample has been exhausted.[34]

An examination of the raw data associated with RaTG13 immediately identified serious anomalies, bringing into question the existence of RaTG13 as a biological entity of completely nature origin.

[34] Dr. Shi Science interview July 2020

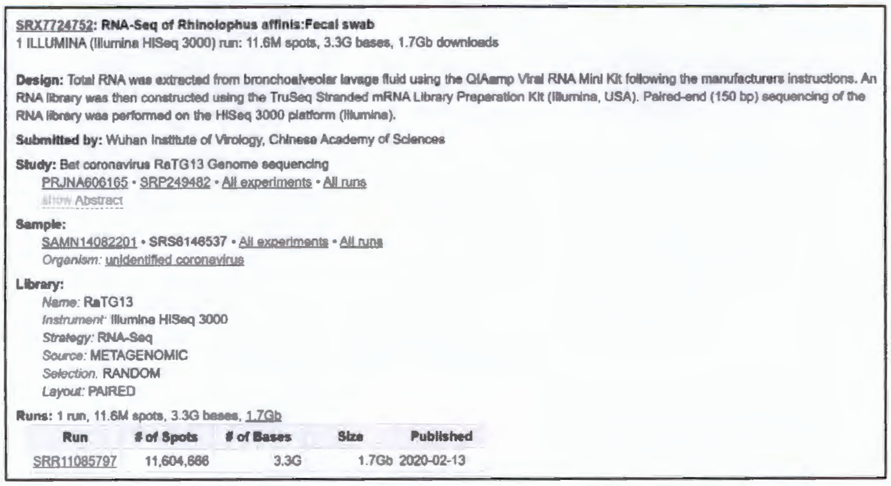
Materials and Methods.

GenBank accession URL table for sequences used in this paper.

The GenBank accession URLs for the specimens, raw reads, and sequences that are used in this paper are contained in the following Table, which can be used to reach the raw data.

Descriptor	URL Hyperlink
SARS-CoV-2 reference sequence in GenBank	SARS-CoV-2 complete genome
Bat coronavirus RaTG13, complete genome, Genbank	RaTG13 complete genome
RaTG13 purported bat fecal specimen	SRR11085797
Rhinolophus bat coronavirus BtCoV/4991 RNA-dependent RNA polymerase (RdRp) gene, partial cds	BtCoV/4991 RdRp gene
SRX8357956: amplicon_sequences of RaTG13	Specimen descriptor
RNA-Seq data for RaTG13	RNA-Seq data for RaTG13
Reference fecal bat specimens from WIV	SRR11085736
Reference fecal bat specimens from WIV	SRR11085734
Reference fecal bat specimens from WIV	SRR11085737
Reference fecal bat specimens from WIV	SRR11085733
Reference fecal bat specimens from WIV	SRR11085735
Reference fecal bat specimens from WIV	SRR11085738
Reference fecal bat specimens from WIV	SRR11085739
Reference fecal bat specimens from WIV	SRR11085740
Reference fecal bat specimens from WIV	SRR11085741

Below is a screen shot of the GenBank entry for the purported specimen from which RaTG13 was identified and upon which RNA-Seq was performed. While the title claims it is a "Rhinolophus affinis fecal swab" specimen it also records in the design of work entry that "(t)otal RNA was extracted from bronchoalveolar lavage fluid." These descriptions are clearly inconsistent.



SRX7724752: **RNA-Seq of Rhinolophus affinis:Fecal swab**
1 ILLUMINA (Illumina HiSeq 3000) run: 11.6M spots, 3.3G bases, 1.7Gb downloads

Design: Total RNA was extracted from bronchoalveolar lavage fluid using the QIAamp Viral RNA Mini Kit following the manufacturers instructions. An RNA library was then constructed using the TruSeq Stranded mRNA Library Preparation Kit (Illumina, USA). Paired-end (150 bp) sequencing of the RNA library was performed on the HiSeq 3000 platform (Illumina).

Submitted by: Wuhan Institute of Virology, Chinese Academy of Sciences

Study: Bat coronavirus RaTG13 Genome sequencing
PRJNA606165 • SRP249482 • All experiments • All runs
Show Abstract

Sample:
SAMN14082201 • SRS6146537 • All experiments • All runs
Organism: unidentified coronavirus

Library:
Name: RaTG13
Instrument: Illumina HiSeq 3000
Strategy: RNA-Seq
Source: METAGENOMIC
Selection: RANDOM
Layout: PAIRED

Runs: 1 run, 11.6M spots, 3.3G bases, 1.7Gb

Run	# of Spots	# of Bases	Size	Published
SRR11085797	11,604,666	3.3G	1.7Gb	2020-02-13

Apparent missing amplicon reads for RaTG13 in GenBank.

There are 33 amplicon reads in GenBank for RaTG13 from experiments recorded as having been performed in 2017 and 2018. A file naming pattern was noticed among the data sets which suggests there may be amplicon runs that were not deposited in GenBank. These files, if related to RaTG13, may contain useful sequence data and an effort should be made to retrieve them and, if appropriate, upload them to GenBank. A Table with the apparently missing data (yellow) is shown here.

Date	Amplicon file name endings							
3-Jun-17	A07	A08						
17-Jun-17	A05	A06						
20-Jun-17						F03	G03	H03
27-Sep-18	A06	B06	C06		E05	F05	G05/G06	H05/H06
29-Sep-18				D05	E05		G04	H04
30-Sep-18	A02	B11						
8-Oct-18			C11				G10	H11
11-Oct-18	A12	B12						
14-Oct-18	A02	B02	C02	D02				

Relationship of *Rhinolophus* bat coronavirus BtCoV/4991 and Bat coronavirus RaTG13.

The Wuhan laboratory has reported on the bat coronaviruses, BtCoV/4991 and RaTG13, in two peer-reviewed publications, one in 2016 and one in February 2020.[35] They have submitted three entries to GenBank for these two viruses, in 2016, February 2020, and May 2020.[36] The GenBank entries confirm sequencing experiments using Sanger dideoxy sequencing in 2016, PCR-generated amplicon sequencing performed on an AB 310 Genetic Analyzer in 2017 and 2018, and RNA-seq performed on an Illumina HiSeq 3000 (instrument id ST-J00123) in 2020. A single GISAID entry records that the RNA-seq data was obtained from an original specimen without passage.[37] This is an important detail since evidence of primate sequences, consistent with VERO cell contamination, is found in this specimen, as reported below, which would suggest laboratory passage.

None of these disclosures report that BtCoV/4991 and RaTG13 are the same coronavirus, simply renamed. This information was only disclosed in a written Question and Answer publication from *Science* magazine by Dr. Shi on July 31, 2020.[4, 38] Given this disclosure months after the original publication concerning RaTG13 in *Nature* it is possible that the omission of the original publication and sequence data concerning BtCoV/4991 violated the "Reporting

[35] 2016 Virologica Sinica paper and February 2020 Nature paper

[36] RaTG13 complete genome Feb 2020, Raw sequence reads for RaTG13 published Feb 2020, Amplicon reads for RaTG13 from 2017 and 2018 published in May 2020.

[37] The GISAID entry is EPI_ISL_402131.

[38] Dr. Shi wrote: "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

standards and availability of data, materials, code and protocols" required for *Nature* publications.[39]

The February 2020 papers uses the RNA-Seq data for RaTG13 genome determination but fails to disclose the previous data obtained by Sanger dideoxy sequencing in 2016 and by amplicon sequencing in 2017 and 2018. Since these unrecorded data establish method-dependent sequencing differences of up to 4% the failure to disclose this data or to reconcile these differences is troubling.

In addition, the raw assembly accession data for RaTG13 are not described or linked to the Genbank entry, MN669532, and also no assembly method is specified in the raw data SRX7724752 12 and the Illumina run. And the amplicon sequencing data has sequence gaps of approximately 20% of the genome. Therefore, no primary assembly data has been made available by the WIV for the RaTG13 genome. This is contrary to the *Nature* Reporting Standards[9] as they state: "When publishing reference genomes, the assembly must be made available in addition to the sequence reads."

Relationship of RaTG13 and SARS-CoV-2.

There have been two descriptions of the process by which the RaTG13 genome was identified as closely homologous to SARS-CoV-2. These seem to be inconsistent with each other.

In the February 2020 *Nature* paper[5] it states:

"We then found that a short region of RNA-dependent RNA polymerase (RdRp) from a bat coronavirus (BatCoV RaTG13)—which was previously detected in Rhinolophus affinis from Yunnan province—showed high sequence identity to 2019-nCoV. We carried out full-length sequencing on this RNA sample (GISAID accession number EPI_ISL_402131). Simplot analysis showed that 2019-nCoV was highly similar throughout the genome to RaTG13, with an overall genome sequence identity of 96.2%."

In a July 2020 interview the process was described:

"We detected the virus by pan-coronavirus RT-PCR in a bat fecal sample collected from Tongguan town, Mojiang county in Yunnan province in 2013, and obtained its partial RdRp sequence. Because the low similarity of this virus to SARS-CoV, we did not pay special attention to this sequence. In 2018, as the NGS sequencing technology and capability in our lab was improved, we did further sequencing of the virus using our remaining samples, and obtained the full-length genome sequence of RaTG13 except the 15 nucleotides at the 5' end. As the sample was used many times for the purpose of viral nucleic acid extraction, there was no more sample after we finished genome sequencing, and we did not do virus isolation and other studies on it. Among all the bat samples we collected, the RaTG13 virus was detected in only one single sample. In 2020, we compared the sequence of SARS-CoV-2 and our unpublished bat

[39] Nature research reporting standards for availability of data

coronavirus sequences and found it shared a 96.2% identity with RaTG13. RaTG13 has never been isolated or cultured."

If the full-length genome of RaTG13 was available by 2018 it is unclear why a database search within the WIV for coronaviruses that resembled SARS-CoV-2 would lead to identifying the 370-nt segment representing the RdRp gene (as stated in the February paper) but not the full length RaTG13 genome (which was stated to have been sequenced by 2018). In addition, an assembly of all available amplicon data for RaTG13 from 2017 and 2018 contains gaps of approximately 20% of the genome. If the sample was completely consumed during the 2017-8 sequencing it is unclear how RNA-Seq was conducted in 2020 to permit the full-length genome to be determined.

Analytical methods. Taxonomy of specimens was determined in the NCBI Sequence Read Archive and KRONA.[40] Blast was used for sequence alignment and comparisons.[41]

To evaluate the data from the bat species relative to the RaTG13 fecal sample analysis, the latter was treated as a fixed result with the comparison to the taxonomy results of the nine bat feces specimens. It also was noted that the data were clearly right skewed (and descriptively both mean/median and standard deviation/interquartile range were used). Therefore, a non-parametric procedure, the Wilcoxon signed-rank test was used with the p-value calculated by an exact procedure because of the small sample size. Considering the synonymous to non-synonymous mutation frequency and how to evaluate that for the various protein coding regions of the virus, it was noted that for all of the genes pooled, the ratio of the synonymous to non-synonymous regions was approximately 0.83. To analyze the corresponding distribution for each gene, we assumed that each mutation was an independent observation from a Bernoulli random variable and, therefore the number of synonymous mutations in the gene would have a binomial distribution (with probability 0.83). A probability was then computed for the actual number of synonymous mutations on this basis (the probability was determined on a one-sided basis, i.e. excess mutations, and was calculated as a strict inequality).

Results.

Original characterization of RaBtCoV/4991 (RaTG13) and related bat fecal specimen.

In 2016 Dr. Shi and colleagues published a paper entitled, "Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft[42]" in which a number of novel bat coronaviruses were isolated from bat fecal specimens collected during 2012 and 2013. The viruses were named, according to the paper, in the following fashion:

[40] NCBI Sequence Archive
[41] Blast alignment
[42] Xing-Yi Ge, et. al., Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft, Virologica Sinica, 2016, 31 (1): 31–40. DOI: 10.1007/s12250-016-3713-9

> "The positive samples detected in this study were named using the abbreviated bat species name plus the bat sample number abbreviation. For example, a virus detected from *Rhinolophus sinicus* in sample number 4017 was named RsBtCoV/4017. If the bat was co-infected by two different coronaviruses, numbers were appended to the sample names, such as RsBtCoV/4017-1 and RsBtCoV/4017-2."

In the July 2020 interview Dr. Shi wrote:

> "Ra4991 is the ID for a bat sample while RaTG13 is the ID for the coronavirus detected in the sample. We changed the name as we wanted it to reflect the time and location for the sample collection. 13 means it was collected in 2013, and TG is the abbreviation of Tongguan town, the location where the sample was collected."

The 2016 and 2020 statements about the naming of virus RsBtCoV/4991 appear inconsistent with each other.

Of the 152 coronaviruses identified, 150 were classified as alphacoronaviruses while only two were classified as betacoronaviruses, HiBtCoV/3740-2 and RaBtCoV/4991. The naming convention from the paper means this latter coronavirus was identified in a fecal specimen from a *Rhinolophus affinis* bat and was sample number 4991.

The latter virus was described in the paper as follows:

> "Virus RaBtCoV/4991 was detected in a R. affinis sample and was related to SL-CoV. The conserved 440-bp RdRp fragment of RaBtCoV/4991 had 89% nt identity and 95% aa identity with SL-CoV Rs672. In the phylogenetic tree, RaBtCoV/4991 showed more divergence from human SARS-CoV than other bat SL-CoVs and could be considered as a new strain of this virus lineage."

The Genbank accession number for RaBtCoV/4991 is MN KP876546.1 and in Genbank it is identified as having been collected in July 2013 as a "feces/swabs" specimen.

The RATG13 genome sequence was assembled from low coverage RNA-Seq data.

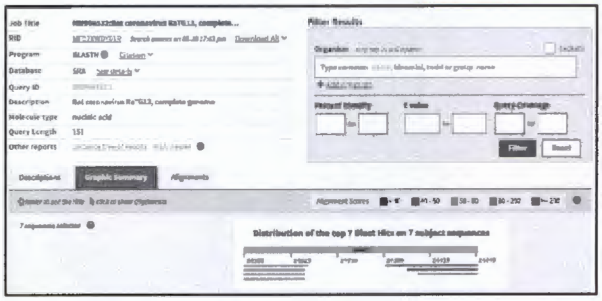
A Blast analysis of the RaTG13 genome against SRR11085797 retrieved about 1700 reads which covers only about 252,000 nt of the total reads of 3.3 Gb. Since the genome size of RaTG13 is known to be about 30,000 nt this represents an 8-fold coverage, typically insufficient for a definitive assembly. For example, some have suggested a 30-fold coverage is necessary to create high quality assemblies.[43]

[43] Sims, D. *et al.* Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews – Genetics. (2014) 15: 121-132. doi:10.1038/nrg3642.

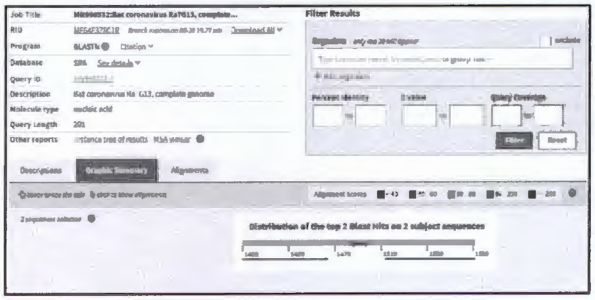
At an eight-fold coverage and based on the typical practice of having four or more reads to call a SNP,[44] the 8-fold coverage of RaTG13 would have 4.2% bases or about 1260 calls of less than 4 reads and about 10 bases would be missed completely, with no calls at all.

A Blast of the RaTG13 published genome onto the RNA-Seq data documents at least two 60 base-pair gaps with no coverage, precluding a complete assembly.

Given the low coverage in the RNA-Seq data, an exploratory, non-exhaustive Blast search was conducted against the published RaTG13 sequence. Two gaps of over 60 nt, shown below, were easily found:



Job Title MN996532:Bat coronavirus RaTG13, complete...
RID [illegible] Download All
Program BLASTN Citation
Database SRA See details
Query ID [illegible]
Description Bat coronavirus RaTG13, complete genome
Molecule type nucleic acid
Query Length 151
Other reports [illegible]



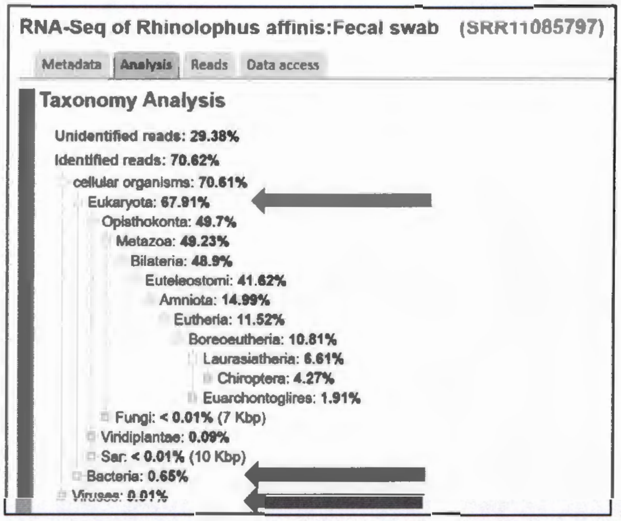
Job Title MN996532:Bat coronavirus RaTG13, complete...
RID [illegible] Download All
Program BLASTN Citation
Database SRA See details
Query ID [illegible]
Description Bat coronavirus RaTG13, complete genome
Molecule type nucleic acid
Query Length 201
Other reports Distance tree of results MSA viewer

It is conceivable there are additional gaps but the above two are sufficient to document that the complete RaTG13 genome sequence could not have been assembled solely from the RNA-Seq data, as stated.[2]

Taxonomy analysis of the RaTG13 specimen is inconsistent with being from bat feces and shows evidence of laboratory cell culture contamination.

According to the Wuhan laboratory, the RaTG13 coronavirus was a fecal swab specimen collected from a *Rhinolophus affinis* bat in 2013. Unexpectedly, (Text-Figure below) the taxonomy analysis is primarily eukaryotic (green arrow; 67.91%) with only traces of bacteria (blue arrow; 0.65%). The viral genomes also make only a trace contribution (red arrow; 0.01%):

[44] Illumina Technical Bulletin Call Coverage



RNA-Seq of Rhinolophus affinis:Fecal swab (SRR11085797)

Taxonomy Analysis

```
Unidentified reads: 29.38%
Identified reads: 70.62%
  cellular organisms: 70.61%
    Eukaryota: 67.91%
      Opisthokonta: 49.7%
        Metazoa: 49.23%
          Bilateria: 48.9%
            Euteleostomi: 41.62%
              Amniota: 14.99%
                Eutheria: 11.52%
                  Boreoeutheria: 10.81%
                    Laurasiatheria: 6.61%
                      Chiroptera: 4.27%
                    Euarchontoglires: 1.91%
        Fungi: < 0.01% (7 Kbp)
      Viridiplantae: 0.09%
      Sar: < 0.01% (10 Kbp)
    Bacteria: 0.65%
  Viruses: 0.01%
```

Taxonomy analysis for RaTG13 data SRR11085797

To compare this specimen composition to bat fecal specimens collected by Dr. Shi and her WIV colleagues and analyzed in other studies, a paper from Dr. Shi's laboratory, also published in February 2020, was identified. In this paper, entitled, "Discovery of Bat Coronaviruses through Surveillance and Probe Capture-Based Next-Generation Sequencing,"[45] a total of nine specimens "collected during previous bat CoV surveillance projects, (were) extracted from bat rectal swabs." According to the Methods section in this paper, the "previous bat CoV surveillance projects" include the field work in 2013 when the RaTG13 was said to have been collected. The comparison below is thus the same specimens collected on the same field surveillance projects by the same investigators from the Wuhan laboratory and sequenced on the same Illumina instrument. These nine specimens will be referred to as "reference fecal specimens" henceforth.

The following Text-Table compares the taxonomical analysis of the RaTG13 and reference fecal specimens. The reference fecal specimens have an average eukaryotic genome content of about 12% while RaTG13's eukaryotic content was 68%. On the other hand, the most abundant genes in the reference fecal specimens were bacterial, with an average of 65%; RaTG13 had less than 1% bacterial genes. And finally, the reference fecal specimens had 1.57% virus genes compared to the 0.01% virus genes of RaTG13.

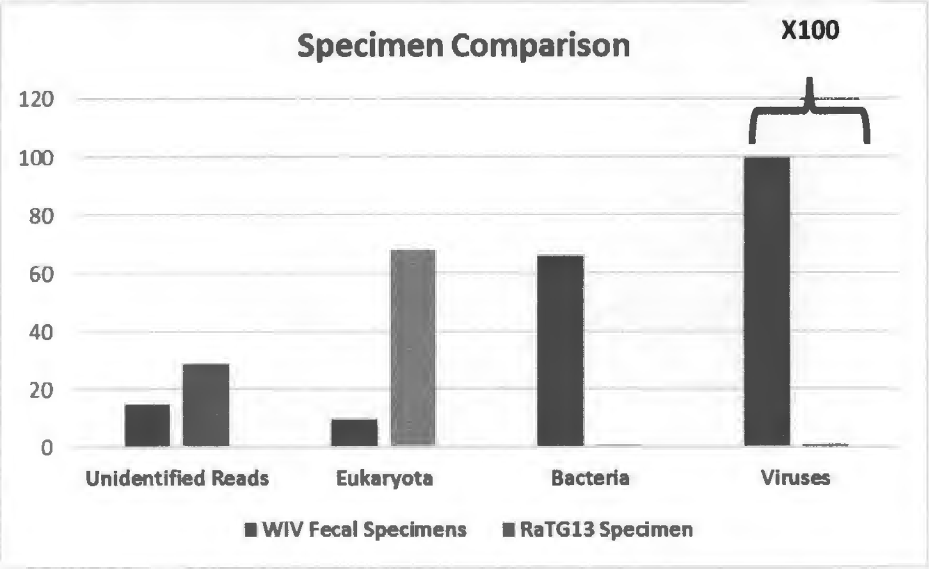
[45] Discovery of bat coronaviruses through surveillance and probe capture-based next-generation sequencing

Specimen ID	Specimen Type	Unidentified Reads	Eukaryota	Bacteria	Viruses	Sum
SRR11085736	*Rhinolophus affinis*	0.86	4.36	91.07	0.03	96.32
SRR11085734	*Miniopterus schreibersii*	3.81	16.03	76.15	0.11	96.1
SRR11085737	*Scotophilus kuhlii*	17.98	8.59	67.81	2.19	96.6
SRR11085733	*Hipposideros larvatus*	13.27	27.99	42.96	4.1	88.32
SRR11085735	*Hipposideros pomona*	34.33	7.96	54.78	0.71	97.78
SRR11085738	*Pipistrellus abramus*	20.33	21.44	47.3	6.45	95.52
SRR11085739	*Tylonycteris pachypus*	61.75	14.34	20.06	0.06	96.21
SRR11085740	*Miniopterus pusillus*	0.78	1.46	99.22	0.05	101.51
SRR11085741	*Rousettus aegyptiacus*	6.44	2.59	88.36	0.45	97.84
Mean +/- SD	Nine bat feces specimens	17.73+/-19.79	11.64+/-9.02	65.30+/-26.10	1.57+/-2.28	96.24+/-3.45
Median +/- IQR	Nine bat feces specimens	13.27+/-24.995	8.59+/-15.26	67.81+/-41.58	0.45+/-3.09	96.32+/-2.00
SRR11085797	**RaTG13 fecal specimen**	**29.38**	**67.91**	**0.65**	**0.01**	**97.95**
	P-value (exact Wilcoxon signed-rank test)	0.16	**0.0039**	**0.0048**	**0.0039**	0.098

As shown in the Text-Table above the RaTG13 specimen is significantly different from the reference fecal specimens in composition. The probabilities for each category, eukaryote, bacteria, and virus, are individually highly statistically significant. They are also independent of each other and therefore the overall probability that RaTG13 has the composition of eukaryote, bacteria, and virus genes that was reported by the Wuhan laboratory but is actually from an authentic bat fecal specimen is less than one in 13 million.

The alternative conclusion is that this sample was not a fecal specimen but was contrived. The data cannot, however, distinguish between a non-fecal specimen that came from true field work on the one hand and a specimen created *de novo* in the laboratory on the other hand.

A graphical comparison of the above data is shown below and visually shows the significant differences between the WIV fecal specimens and the RaTG13 specimen, despite the claim they were collected in the same field surveillance trips:



Another comparison can be made between the reference fecal specimens and the RaTG13 specimen by looking at the taxonomy of the nine to twelve "strong signals" identified on the NCBI Sequence Read Archive. The following Text-Table is a summary of these findings.

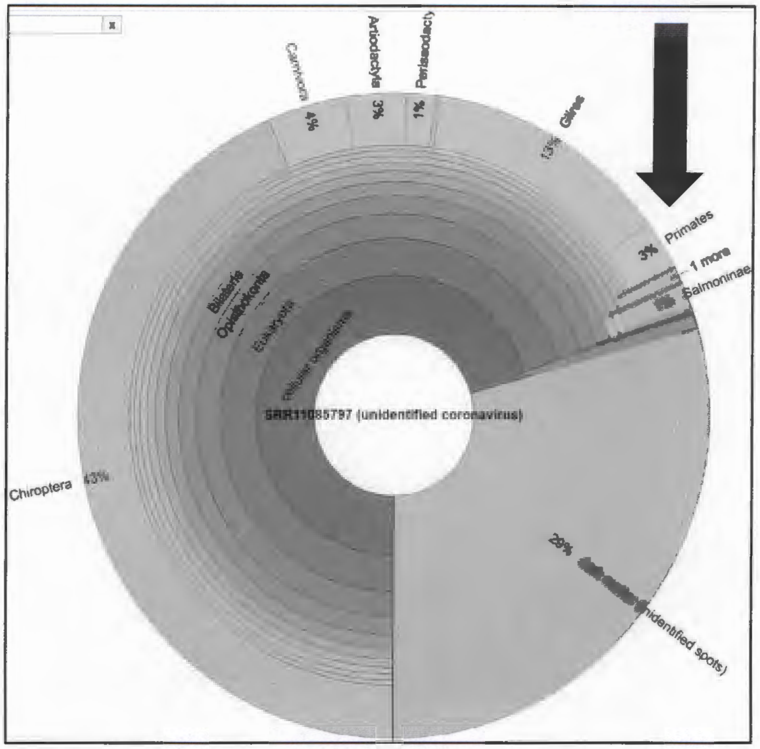
Specimen	The identity of the Strong Signals in the Specimens		
	Bacteria	Eukaryotes	Viruses
Rhinolophus affinis anal swab (SRR11085736)	92%	One magnaorder of placental mammals, includes bat	None
Miniopterus schreibersii anal swab (SRR11085734)	88%	**One bat**, the host bat, Miniopterus sp.	None
Scotophilus kuhlii anal swab (SRR11085737)	56%	**Two bats**, mouse-eared and big brown bats.	Two viruses, kobuvirus (host includes bats) and a Scotophilus kuhlii coronavirus
Hipposideros larvatus anal swab (SRR11085733)	56%	**One bat,** the host bat, Hipposideros sp. and one rodent.	Hipposideros pomona bat coronavirus
Hipposideros pomona: Anal swab (SRR11085735)	78%	**One bat**, the host bat, Hipposideros sp.	None
Pipistrellus abramus: Anal swab (SRR11085738)	73%	**Two bats**, the big brown bat and the mouse-eared bat.	Pipistrellus abramus bat coronavirus
Tylonycteris pachypus: Anal swab (SRR11085739)	67%	**Three bats**, the microbat, the great roundleaf bat, and a superorder of mammals, which includes bats.	None
Miniopterus pusillus: Anal swab (SRR11085740)	89%	**One bat**, the Natal long-fingered bat.	None
Rousettus aegyptiacus: Anal swab (SRR11085741)	91%	One magnaorder of placental mammals, includes bats.	None
Average	77%		
RaTG13 Rhinolophus affinis:Fecal swab (SRR11085797)	None	All nine strong signals are eukaryotes. **Five bats**, the Great Roundleaf bat, resident of China, the Egyptian fruit bat, which is not found in China, a megabat, mouse-eared bat, and bent-winged bat. Two marmots, the Alpine marmot from Europe and the Yellow-bellied marmot of North America.The paraorder of whales. The red fox.	None

As can be seen, while the strong signals in the authentic specimens contain 56% to 92% (average 77%) bacterial signals, the RaTG13 specimen has no bacteria among the nine strong signals. Most specimens do not have virus strong signals but the three that do are host-related coronaviruses (four) or one host-related kobuvirus.

RaTG13 has no viral strong signals. Among the reference specimens with eukaryotic strong signals, they are either bat-related genes (eleven) or higher order taxonomy signals that include bats (three). There is one anomalous rodent-related signal among the reference specimens.

The RaTG13 specimen is again an outlier with all nine strong signals arising from eukaryotic genes. Five of the nine signals are bats, some resident to China and some with non-Chinese host ranges. Surprisingly, unlike three of the reference bat signals which are identified as host-related, the RaTG13 specimen did not contain *Rhinolophus* sp. host-related strong signals. The remaining four strong signals are marmot-related genes (two), whale-related gene (one), and red fox-related gene (one).

Finally, a Krona analysis (below) identifies 3% primate sequences (red arrow) in the RaTG13 sequence data. This is consistent with contamination by the standard laboratory coronavirus cell culture system, the VERO monkey kidney cell line.



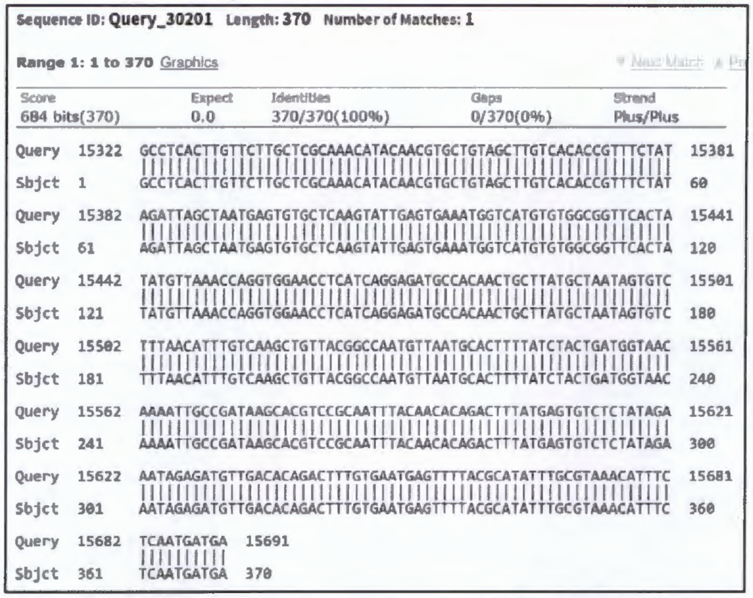
Source: Krona analysis of RaTG13 specimen

It is unclear why these obviously anomalous findings were not detected during the peer-review process prior to publication of this important work. At this point, an explanation is needed from the WIV to refute the conclusion that the specimen identified as the source of RaTG13 is **not** a bat fecal/anal specimen and that the primate genetic material is consistent with a VERO cell contaminated specimen.

Method-related nt base substitutions in RaTG13.

The original Sanger dideoxy RdRp sequence reported in 2016 is homologous to RNA-seq data from 2020 but is non-homologous to amplicon sequencing data from 2017 and 2018.

As expected, a comparison of the 2016 RdRp GenBank sequence for BtCoV/4991 obtained by Sanger dideoxy sequencing with the RNA-seq sequencing of RaTG13 reported in *Nature* shows 100% identity over the 370 nt segment.



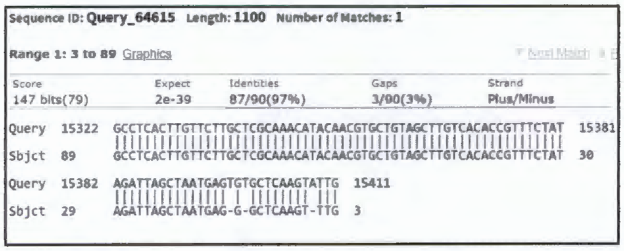
Sequence ID: Query_30201 Length: 370 Number of Matches: 1

Range 1: 1 to 370 Graphics

Score	Expect	Identities	Gaps	Strand
684 bits(370)	0.0	370/370(100%)	0/370(0%)	Plus/Plus

```
Query  15322  GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  15381
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1      GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  60

Query  15382  AGATTAGCTAATGAGTGTGCTCAAGTATTGAGTGAAATGGTCATGTGTGGCGGTTCACTA  15441
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61     AGATTAGCTAATGAGTGTGCTCAAGTATTGAGTGAAATGGTCATGTGTGGCGGTTCACTA  120

Query  15442  TATGTTAAACCAGGTGGAACCTCATCAGGAGATGCCACAACTGCTTATGCTAATAGTGTC  15501
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121    TATGTTAAACCAGGTGGAACCTCATCAGGAGATGCCACAACTGCTTATGCTAATAGTGTC  180

Query  15502  TTTAACATTTGTCAAGCTGTTACGGCCAATGTTAATGCACTTTTATCTACTGATGGTAAC  15561
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181    TTTAACATTTGTCAAGCTGTTACGGCCAATGTTAATGCACTTTTATCTACTGATGGTAAC  240

Query  15562  AAAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGA  15621
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241    AAAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGA  300

Query  15622  AATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  15681
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301    AATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTC  360

Query  15682  TCAATGATGA  15691
              ||||||||||
Sbjct  361    TCAATGATGA  370
```

Surprisingly, the two amplicon sequences from 2017 that partially cover the 370 nt RdRp region have four base substitutions or gaps over a total segment of 219 nt (2% divergence).

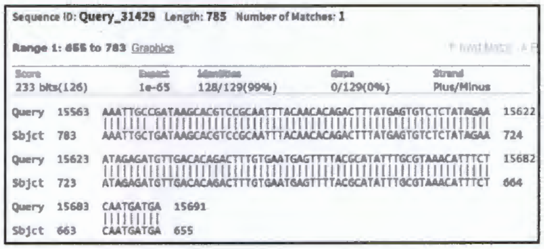


Sequence ID: Query_64615 Length: 1100 Number of Matches: 1

Range 1: 3 to 89 Graphics

Score	Expect	Identities	Gaps	Strand
147 bits(79)	2e-39	87/90(97%)	3/90(3%)	Plus/Minus

```
Query  15322  GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  15381
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  89     GCCTCACTTGTTCTTGCTCGCAAACATACAACGTGCTGTAGCTTGTCACACCGTTTCTAT  30

Query  15382  AGATTAGCTAATGAGTGTGCTCAAGTATTG  15411
              ||||||||||||||| | |||||||| |||
Sbjct  29     AGATTAGCTAATGAG-G-GCTCAAGT-TTG  3
```



Sequence ID: Query_31429 Length: 785 Number of Matches: 1

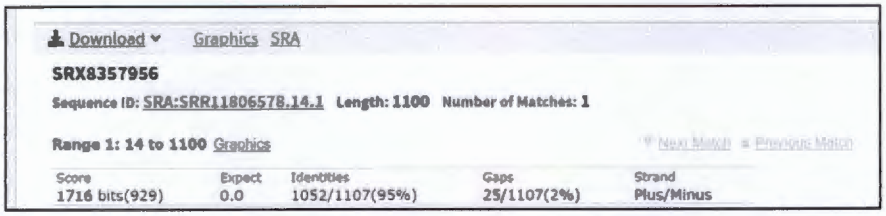
Range 1: 655 to 783 Graphics

Score	Expect	Identities	Gaps	Strand
233 bits(126)	1e-65	128/129(99%)	0/129(0%)	Plus/Minus

```
Query  15563  AAATTGCCGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGAA  15622
              ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  783    AAATTGCTGATAAGCACGTCCGCAATTTACAACACAGACTTTATGAGTGTCTCTATAGAA  724

Query  15623  ATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTCT  15682
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  723    ATAGAGATGTTGACACAGACTTTGTGAATGAGTTTTACGCATATTTGCGTAAACATTTCT  664

Query  15683  CAATGATGA  15691
              |||||||||
Sbjct  663    CAATGATGA  655
```

RaTG13 Spike Protein gene has 5% substitutions when comparing 2020 RNA-Seq and 2017 amplicon sequencing data.

The segment of RaTG13 which shows the greatest sequence divergence between the RNA-seq and amplicon sequencing methods spans from A8886 to A9987 and is shown here below. It contains 80 base substitutions/indels in a 1107 nt sequence (5% substitution and 2% gaps).



SRX8357956

Sequence ID: SRA:SRR11806578.14.1 Length: 1100 Number of Matches: 1

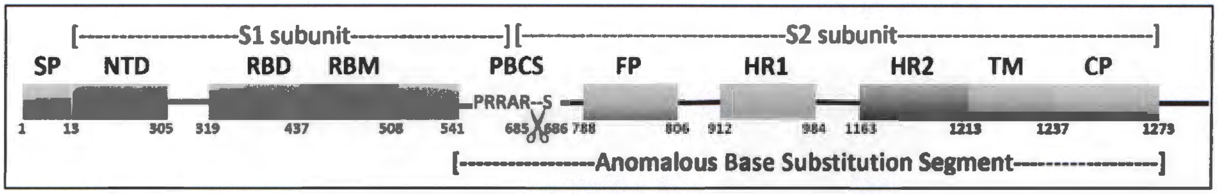
Range 1: 14 to 1100 Graphics

Score	Expect	Identities	Gaps	Strand
1716 bits(929)	0.0	1052/1107(95%)	25/1107(2%)	Plus/Minus

No explanation has been offered in publications from the WIV for the method-dependent sequencing differences identified here, which are twenty- to 50-fold higher than the 0.1% technical error rate sometimes attributed to RNA-Seq data.

The Spike Protein gene sequence substitution divergence between RaTG13 and SARS-CoV-2 contains an improbable synonymous/non-synonymous pattern.

The functional structure of the SARS-CoV-2 Spike Protein is shown here:


[-----------------S1 subunit-----------------][-----------------S2 subunit-----------------]
SP
NTD
RBD
RBM
PBCS
FP
HR1
HR2
TM
CP
PRRAR--S
1
13
305
319
437
508
541
685
686
788
806
912
984
1163
1213
1237
1273
[---------------Anomalous Base Substitution Segment---------------]

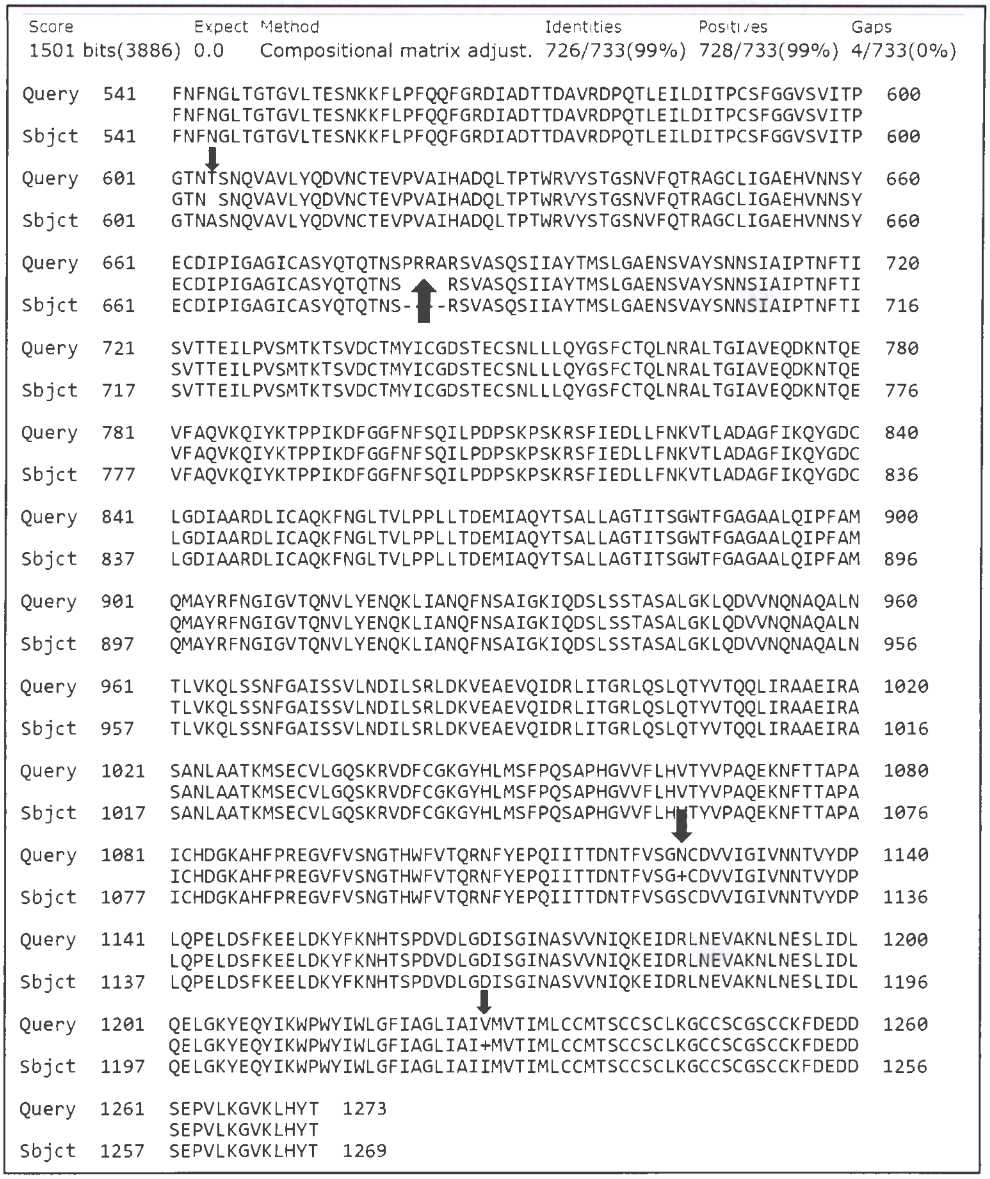
The SARS-CoV-2 Spike protein (above) contains an S1 subunit and S2 subunit with the Polybasic Cleavage Site (PBCS) between R685 and S686. This cleavage is performed by a host cell surface protease, furin, and is an important attribute in explaining the virulence of SARS-CoV-2 compared to other human coronaviruses, which do not have a furin cleavage site. The PBCS also contains the unusual PRRA insertion that has not been previously seen in Clade B coronaviruses and for which no natural mechanism for its appearance has been offered.[46]

The S1 subunit is located within the N-terminal 14–685 amino acids of S protein, containing N-terminal domain (NTD), receptor binding domain (RBD), and receptor binding motif (RBM). The S2 subunit contains a fusion peptide (FP), heptad repeat 1 (HR1), heptad repeat 2 (HR2), transmembrane domain (TM) and cytoplasmic domain (CP).

The base substitution pattern of synonymous and non-synonymous substitutions when comparing RaTG13 and the reference sequence of SARS-CoV-2 demonstrated an anomalous pattern for the coding region for aa 541 to 1273, a 733 aa protein segment representing over 60% of the SP gene.

As shown in the Text-Figure below, there are only three substitutions (red arrow) and the PBCS insertion (blue arrow) when comparing this segment of the RaTG13 and SARS-CoV-2 SP. Excluding the PBCS, the amino acid sequences are 99.6% identical.

[46] The proximal origin of SARS-CoV-2.



```
 Score              Expect  Method                        Identities      Positives       Gaps
 1501 bits(3886)  0.0     Compositional matrix adjust.  726/733(99%)  728/733(99%)  4/733(0%)

Query  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600
             FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
Sbjct  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600

Query  601   GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
             GTN SNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
Sbjct  601   GTNASNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660

Query  661   ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
             ECDIPIGAGICASYQTQTNS    RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
Sbjct  661   ECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716

Query  721   SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
             SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
Sbjct  717   SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  776

Query  781   VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  840
             VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
Sbjct  777   VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  836

Query  841   LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  900
             LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM
Sbjct  837   LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM  896

Query  901   QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  960
             QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN
Sbjct  897   QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN  956

Query  961   TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1020
             TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
Sbjct  957   TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA  1016

Query  1021  SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1080
             SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
Sbjct  1017  SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA  1076

Query  1081  ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP  1140
             ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSG+CDVVIGIVNNTVYDP
Sbjct  1077  ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGSCDVVIGIVNNTVYDP  1136

Query  1141  LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL  1200
             LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
Sbjct  1137  LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL  1196

Query  1201  QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD  1260
             QELGKYEQYIKWPWYIWLGFIAGLIAI+MVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
Sbjct  1197  QELGKYEQYIKWPWYIWLGFIAGLIAIIMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD  1256

Query  1261  SEPVLKGVKLHYT  1273
             SEPVLKGVKLHYT
Sbjct  1257  SEPVLKGVKLHYT  1269
```

Given the high amino acid identity of this 733 amino acid sequence (except for the PBCS insertion) and the typical coronavirus synonymous to non-synonymous mutation frequency of between three and five synonymous mutations for each non-synonymous mutation,[47] it was expected that a comparison of the nucleotide sequence for this region between SARS-CoV-2 and RaTG13 would show an almost identical sequence as well.

In fact, when the SARS-CoV-2 nt sequence 23,183-25,384 was compared to the RaTG13 nt sequence 23,165-25,354, the corresponding genome sequence to the 99.6% identical protein sequence above, the nucleotide identity was only 94.2% identical, with 122 synonymous substitutions and only the three non-synonymous substitutions.

[47] Comparative genomic analysis

To put this in context a comparison of thirteen other protein coding regions of SARS-CoV-2 and RaTG13 (Text-Table below) shows that the overall synonymous to non-synonymous mutation frequency is 549 synonymous to 109 non-synonymous or a ratio of about 5.0.

Gene	Region of Genome	Total Nucleotides	Synonymous mutations	Non-Synonymous mutations	S/NS	Probability of more than the number of synonymous mutations given the probability of a synonymous mutation is 0.83 (based on all genes pooled)
pp1ab	1-21,239	21,239	659	102	6.5	0.003
pp1ab ABSS	**7448-18266**	**10,818**	**283**	**13**	**21.8**	**5.73 x 10^-12**
Spike Protein RBD	1-1814	1814	131	27	4.9	0.48
Anomalous Base Substitution Segment	**23,183-25,384**	**2201**	**112**	**3**	**37.3**	**< 1.0 x 10^-7**
Entire Spike Protein	1-3810	3808	231	41	5.6	0.18
ORF1a polyprotein	1-13,215	13215	440	86	5.2	0.33
ORF3a protein	1-828	828	25	6	4.2	0.56
E Protein	1-228	228	1	0	Infinite	0.83
M Protein	1-669	669	27	3	9.0	0.1
ORF6 Protein	1-186	186	3	0	Infinite	0.17
ORF7a Protein	1-366	366	13	3	4.3	0.47
ORF7b Protein	1-132	132	0	1	0	0.83
ORF8 Protein	1-366	366	5	6	0.8	0.99
Nucleocapsid Phosphoprotein	1-1260	1260	35	4	8.75	0.083

With the exception of the anomalous base substitution segment (ABSS) in the Spike Protein gene and the pp1ab gene, the remainder of the S/SN substitution ratios are consistent with the literature values for coronaviruses. Only two genes or gene regions have a higher S/SN ratio than the ABSS because they have no non-synonymous mutations: the E protein gene with 228 nucleotides and the ORF6 protein gene with 186 nucleotides. Because of the short length of these two genes, the probabilities of the results for the E and ORF6 genes were not significant, with p-values of 0.86 and 0.17, respectively.

The p-value for the ABSS, on the other hand, was highly significant, with a p-value of <0.0000001. This strongly suggests a non-natural cause for this base substitution pattern, barring some unknown biological mechanism for such a result.

A second highly anomalous sequence was found in the pp1ab gene. This is about five-times larger than the Spike Protein region and is even more unlikely to have happened naturally, a chance of about one in 100 billion times.

Are there only synonymous mutations in these regions because non-synonymous mutations lead to non-replicative viruses?

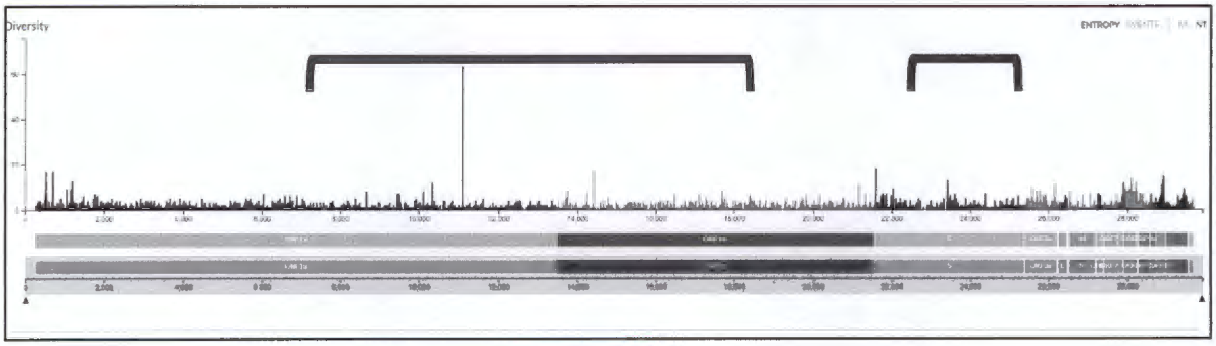
A simple explanation for these results would be an extreme criticality for the specific sequences of these regions with respect to infectivity. If a single amino acid change yielded a non-transmissible viral particle that strong negative purification process could explain the above results.

This hypothesis can be immediately rejected based on two observations.

In an examination of over 80,000 SARS-CoV-2 genome sequences, the most common Spike Protein non-synonymous mutation is within the ABSS (D614G) which was identified within weeks of the outbreak in January 2020 and which has become "the dominant virus...in every geographical region."[48] Specifically, as of August 28, 2020, GISAID reports that 65,738 full length SARS-CoV-2 genomes of a total of 83,387, or 79%, and comprising the G, GH, and GR clades, contain the D614G SNV. Under real world biological conditions, the ABSSN region has in fact, not a strong negative purification process in operation but in fact a strong positive selection process ongoing.

Secondly, in an analysis of mutations in 63,421 SARS-CoV-2 genomes the Spike Protein amino acid 605 to 1120 region had a total of 7,149 mutations. Fully 5,936 of these mutations (83%) are the above noted D614G non-synonymous change. Of the remaining 1213 mutations, 452 were non-synonymous while 755 were synonymous, a ratio of 1.7. There were also four indels and two stop codon mutations.

The following Text-Figure contains a map of the SARS-CoV-2 genome with the location of amino acid changes that have been found during the worldwide spread noted, with the frequency related to the height of the mark. The two ABSS in pp1ab and SP are marked with red brackets and clearly demonstrate an abundance of non-synonymous mutations in these regions during the human-to-human spread.



Nextstrain SARS-CoV-2 amino acid change events

Clearly, these regions can tolerate many non-synonymous mutations, rejecting the theory of a criticality for the amino acid sequence of this region. No other natural biological mechanism to explain these results has been identified.

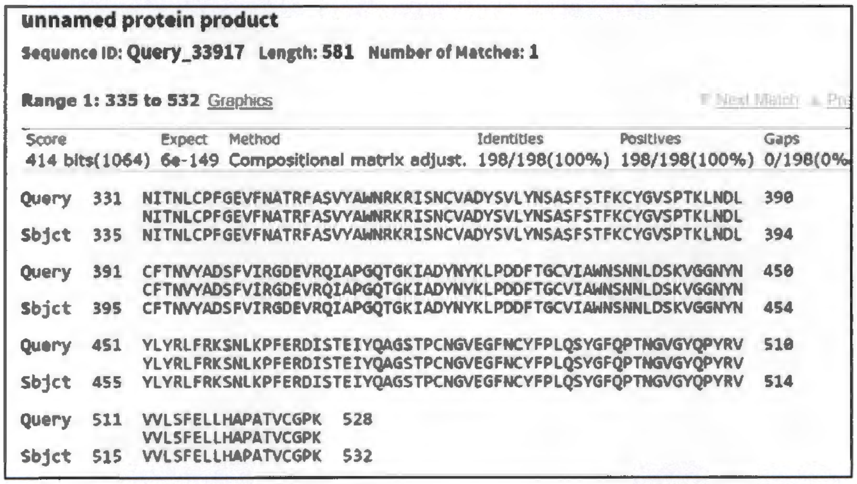
Codon modification, enhancement, or optimization is an example from synthetic biology in which the S/SN ratio is, by design, an anomaly when looked at through the lens of nature

[48] Biswas NK, Majumder PP. Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. Indian J Med Res. 2020;151(5):450-458. doi:10.4103/ijmr.IJMR_1125_20.

Synonymous codon substitution is a decades old, well known method of enhancing gene expression when cloning exogenous genes in a laboratory experiment. In a paper on the immunogenicity of the SARS-CoV-2 Spike Protein[49] the following synthetic biology methods were used:

"We used the following structure coordinates of the coronavirus spike proteins from the PDB to define the boundaries for the design of RBD expression constructs: SARS-CoV-2 (6VSB), SARS-CoV-1 (6CRV), HKU-1 (5I08), OC43 (6NZK), 229E (6U7H) NL63 (6SZS). Accordingly, a codon-optimized gene encoding for S1-RBD [SARS-CoV-1 (318 – 514 aa, P59594), SARS-CoV-2 (331 – 528 aa, QIS60558.1), OC43 (329 – 613 aa, P36334.1), HKU-1 (310 – 611 aa, Q0ZME7.1), 229E (295 – 433 aa, P15423.1) and NL63 (480 – 617 aa, Q6Q1S2.1)] containing human serum albumin secretion signal sequence, three purification tags (6xHistidine tag, Halo tag, and TwinStrep tag) and two TEV protease cleavage sites was cloned into the mammalian expression vector pαH. S1 RBDs were expressed in Expi293 cells (ThermoFisher) and purified from the culture supernatant by nickel-nitrilotriacetic acid agarose (Qiagen)."

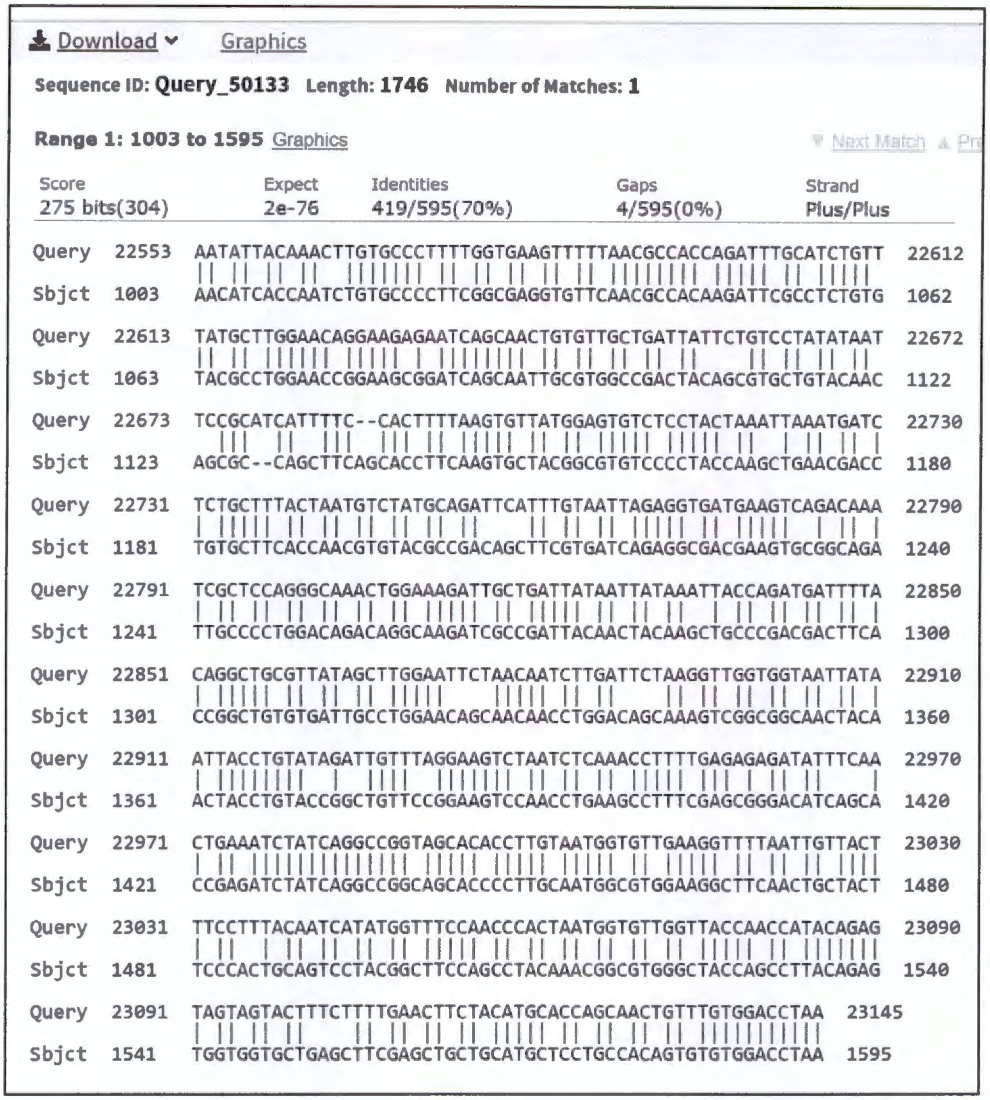
The Genbank alignment (below) confirms that the authentic SARS-CoV-2 Spike Protein sequence (https://www.ncbi.nlm.nih.gov/nuccore/1798174254) and the Synthetic construct SARS CoV-2 spike protein receptor binding domain gene, complete cds are 100% homologous at the protein level:



```
unnamed protein product
Sequence ID: Query_33917  Length: 581  Number of Matches: 1

Range 1: 335 to 532

Score           Expect  Method                        Identities     Positives      Gaps
414 bits(1064)  6e-149  Compositional matrix adjust.  198/198(100%)  198/198(100%)  0/198(0%

Query  331   NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL  390
             NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL
Sbjct  335   NITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDL  394

Query  391   CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN  450
             CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN
Sbjct  395   CFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYN  454

Query  451   YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV  510
             YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV
Sbjct  455   YLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV  514

Query  511   VVLSFELLHAPATVCGPK  528
             VVLSFELLHAPATVCGPK
Sbjct  515   VVLSFELLHAPATVCGPK  532
```

But a comparison of the authentic nucleotide sequence of SARS-CoV-2 to the codon-optimized synthetic construct shows no match using the "highly similar Megablast" algorithm setting. When the alignment algorithm is run in a more relaxed mode the impact of codon optimization in this case can be seen, a 70% homology:

[49] https://immunology.sciencemag.org/content/5/48/eabc8413/tab-pdf



Sequence ID: Query_50133 Length: 1746 Number of Matches: 1

Range 1: 1003 to 1595 Graphics

Score	Expect	Identities	Gaps	Strand
275 bits(304)	2e-76	419/595(70%)	4/595(0%)	Plus/Plus

```
Query  22553  AATATTACAAACTTGTGCCCTTTTGGTGAAGTTTTTAACGCCACCAGATTTGCATCTGTT  22612
              || || || ||  ||||||| || || || || || |||||||| ||||| || ||||| 
Sbjct  1003   AACATCACCAATCTGTGCCCCTTCGGCGAGGTGTTCAACGCCACAAGATTCGCCTCTGTG  1062

Query  22613  TATGCTTGGAACAGGAAGAGAATCAGCAACTGTGTTGCTGATTATTCTGTCCTATATAAT  22672
              || || |||||| ||||| | |||||||| || || || || ||    || || || || 
Sbjct  1063   TACGCCTGGAACCGGAAGCGGATCAGCAATTGCGTGGCCGACTACAGCGTGCTGTACAAC  1122

Query  22673  TCCGCATCATTTTC--CACTTTTAAGTGTTATGGAGTGTCTCCTACTAAATTAAATGATC  22730
                |||  ||  |||  ||| || ||||| || || ||||| ||||| ||  | || || |
Sbjct  1123   AGCGC--CAGCTTCAGCACCTTCAAGTGCTACGGCGTGTCCCCTACCAAGCTGAACGACC  1180

Query  22731  TCTGCTTTACTAATGTCTATGCAGATTCATTTGTAATTAGAGGTGATGAAGTCAGACAAA  22790
              | ||||| || || || || ||    || || || ||||| || |||||  | || |
Sbjct  1181   TGTGCTTCACCAACGTGTACGCCGACAGCTTCGTGATCAGAGGCGACGAAGTGCGGCAGA  1240

Query  22791  TCGCTCCAGGGCAAACTGGAAAGATTGCTGATTATAATTATAAATTACCAGATGATTTTA  22850
              | || || || || || || ||||| || ||||| || || ||  | || || || || |
Sbjct  1241   TTGCCCCTGGACAGACAGGCAAGATCGCCGATTACAACTACAAGCTGCCCGACGACTTCA  1300

Query  22851  CAGGCTGCGTTATAGCTTGGAATTCTAACAATCTTGATTCTAAGGTTGGTGGTAATTATA  22910
              | ||||| || || || |||||    ||||| || ||    || || || || || || |
Sbjct  1301   CCGGCTGTGTGATTGCCTGGAACAGCAACAACCTGGACAGCAAAGTCGGCGGCAACTACA  1360

Query  22911  ATTACCTGTATAGATTGTTTAGGAAGTCTAATCTCAAACCTTTTGAGAGAGATATTTCAA  22970
              | ||||||||  |  ||||  ||||||| || || || ||||| ||| | || ||    |
Sbjct  1361   ACTACCTGTACCGGCTGTTCCGGAAGTCCAACCTGAAGCCTTTCGAGCGGGACATCAGCA  1420

Query  22971  CTGAAATCTATCAGGCCGGTAGCACACCTTGTAATGGTGTTGAAGGTTTTAATTGTTACT  23030
              | || ||||||||||||||| ||||| ||||| ||||| || ||||| || || || ||||
Sbjct  1421   CCGAGATCTATCAGGCCGGCAGCACCCCTTGCAATGGCGTGGAAGGCTTCAACTGCTACT  1480

Query  23031  TTCCTTTACAATCATATGGTTTCCAACCCACTAATGGTGTTGGTTACCAACCATACAGAG  23090
              | ||  | || || || || ||||| || || || || || || ||||| || |||||||
Sbjct  1481   TCCCACTGCAGTCCTACGGCTTCCAGCCTACAAACGGCGTGGGCTACCAGCCTTACAGAG  1540

Query  23091  TAGTAGTACTTTCTTTTGAACTTCTACATGCACCAGCAACTGTTTGTGGACCTAA  23145
              | || || ||    || || || || ||||| || || || || ||||||||||||
Sbjct  1541   TGGTGGTGCTGAGCTTCGAGCTGCTGCATGCTCCTGCCACAGTGTGTGGACCTAA  1595
```

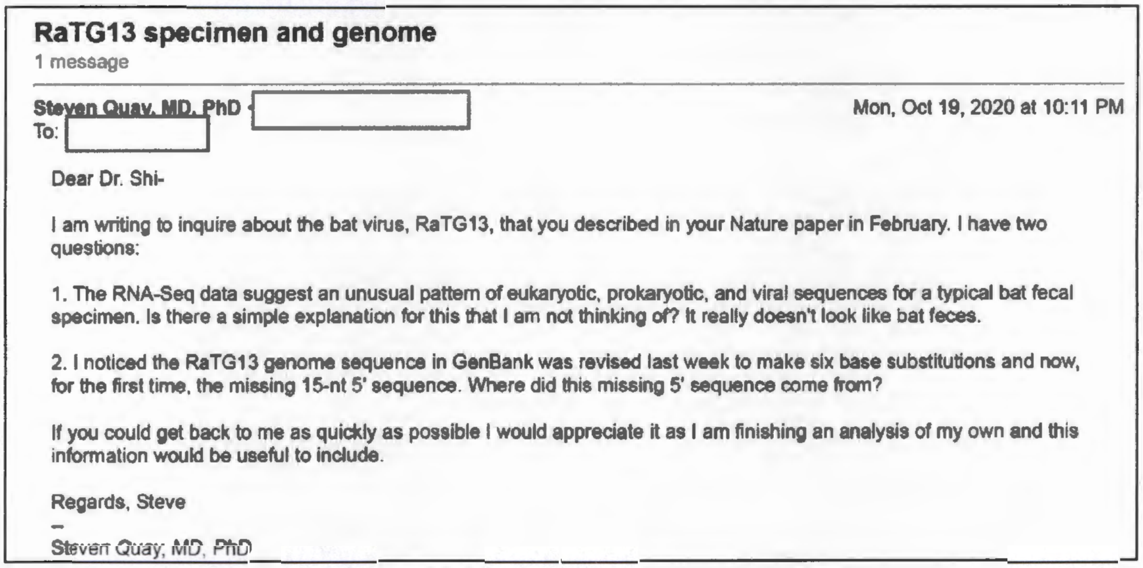
This is a situation in which there are 176 synonymous changes without a single non-synonymous change and is the genome signature of laboratory-derived synthetic biology. If these sequences were compared for phylogenetic divergence without the knowledge of their artificial construction, this synthetic laboratory experiment would create the impression that these two sequences had diverged in the wild from a common ancestor decades earlier.

The following Table identifies four regions of the RaTG13 and SARS-CoV-2 genomes in which there were a total of 220 synonymous mutations without a single non-synonymous change.

Protein/Gene	Protein Region	Total Nucleotides	Synonymous mutations	NS Mutations
S Protein	605-1124	1557	91	0
pp1ab	3607-4534	2781	66	0
pp1ab	4626-5111	1455	26	0
pp1ab	5113-5828	2145	37	0
	Total	7938	220	0

These regions represent over 26% of the entire genome and appear analogous to the outcome expected from the application of a synonymous codon modified, laboratory-derived synthetic biology project. They also represent about one-sixth of the 4% apparent phylogenetic divergence between RaTG13 and SARS-CoV-2.

October GenBank update. On October 13, 2020 the sequence for RaTG13 was updated. For the first time the first 15 nucleotides at the 5' end were present. However, these were not found in a blast of either the RNA-Seq raw reads or the Amplicons. The following email was sent to Dr. Shi asking for an explanation of the fecal specimen composition and the source for the 5' nt data.



RaTG13 specimen and genome
1 message

Steven Quay, MD, PhD [redacted] Mon, Oct 19, 2020 at 10:11 PM
To: [redacted]

Dear Dr. Shi-

I am writing to inquire about the bat virus, RaTG13, that you described in your Nature paper in February. I have two questions:

1. The RNA-Seq data suggest an unusual pattern of eukaryotic, prokaryotic, and viral sequences for a typical bat fecal specimen. Is there a simple explanation for this that I am not thinking of? It really doesn't look like bat feces.

2. I noticed the RaTG13 genome sequence in GenBank was revised last week to make six base substitutions and now, for the first time, the missing 15-nt 5' sequence. Where did this missing 5' sequence come from?

If you could get back to me as quickly as possible I would appreciate it as I am finishing an analysis of my own and this information would be useful to include.

Regards, Steve

--
Steven Quay, MD, PhD

At the time of this writing a response has not been received.

Discussion. The foundation of the working hypothesis that the COVID-19 pandemic arose via a natural zoonotic transfer from a non-human vertebrate host to man has been built on two publications: the February 3, 2020 *Nature* paper by Dr. Zheng-Li Shi and colleagues, in which the bat coronavirus RaTG13 is first identified as the closest sequence identity to SARS-CoV-2 at 96.2% and the March 17, 2020 *Nature Medicine* paper entitled, "The proximal origin of SARS-CoV-2," by Andersen *et al.*, in which the Shi *et al.* paper is cited as evidence for a bat origin for the pandemic. In the approximately six months since they were published, these two papers have been cited over 1600- and 200-times on PubMed, respectively.

However, research is beginning to question whether a bat species can be considered a natural reservoir for SARS-CoV-2. A recent paper performed an *in silico* simulation of the SARS-CoV-2 Spike Protein interaction with the cell surface receptor, ACE2, from 410 unique vertebrate species, including 252 mammals.[50] Among primates, 18/19 have an ACE2 receptor which is

[50] Broad host range of SARS-CoV-2 predicted by comparative and structural analysis of ACE2 in vertebrates Joana Damas, et al. Proc. of the Nat. Acad. of Sci. Aug 2020, 202010146; DOI: 10.1073/pnas.2010146117

100% homologous to the human protein in the 25 residues identified to be critical to infection, including the *Chlorocebus sabaeus* (the Old World African Green monkey) and the rhesus macaques.

It is noteworthy that the laboratory workhorse of coronavirus research is the VERO cell, isolated from a female African Green monkey in 1962, and containing an ACE2 receptor that is 100% homologous to the human ACE2 in the 25 critical amino acids for infectivity.

This *in silico* work was confirmed in the laboratory with respect to rhesus macaques. Within weeks of the identification of SARS-CoV-2, the Wuhan laboratory had demonstrated that the pandemic virus would infect and produce a pneumonia in rhesus macaques.[51]

A surprising finding from the ACE2 *in silico* surveillance work was the very poor predicted affinity of the ACE2 receptors in both bats and pangolins. Of 37 bat species studied, 8 scored low and 29 scored very low. As expected by these predictions, cell lines derived from big brown bat (*Eptesicus fuscus*),[52] Lander's horseshoe bat (*Rhinolophus landeri*), and Daubenton's bat (*Myotis daubentonii*) could not be infected with SARS-CoV-2.[53]

It is unfortunate that growth of the RaTG13 specimen could not have been attempted in the *Rhinolophus sinicus* primary or immortalized cells generated and maintained in the Wuhan laboratory: kidney primary cells (RsKi9409), lung primary cells (RsLu4323), lung immortalized cells (RsLuT), brain immortalized cells (RsBrT) and heart immortalized cells (RsHeT).[54] However it should be noted that a synthetically created RaTG13 was reported not to infect human cells expressing *Rhinolophus sinicus* ACE2, providing evidence that RaTG13 may not be a viable coronavirus in a wild bat population.[55]

The other proposed intermediate host, the pangolin, also had predicted ACE-2 affinity that was either low or very low.

A recent paper that examined the high synonymous mutation difference between RaTG13 and SARS-CoV-2 used an *in silico* methodology to suggest that the difference could be largely attributed to the RNA modification system of hosts.[56] However, the authors do not "(t)he

[51] Infection with Novel Coronavirus (SARS-CoV-2) Causes Pneumonia in the *Rhesus Macaques.* C. Shan et al., Research Square, **DOI:** 10.21203/rs.2.25200/v1. Shan, C., Yao, Y., Yang, X. *et al.* Infection with novel coronavirus (SARS-CoV-2) causes pneumonia in *Rhesus macaques. Cell Res* **30,** 670–677 (2020). https://doi.org/10.1038/s41422-020-0364-z

[52] J. Harcourt et al., Severe acute respiratory syndrome coronavirus 2 from patient with coronavirus disease, United States. Emerg. Infect. Dis. 26, 1266–1273 (2020).

[53] M. Hoffmann et al., SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181, 271–280.e8 (2020).

[54] Zhou, P., Fan, H., Lan, T. et al. Fatal swine acute diarrhoea syndrome caused by an HKU2-related coronavirus of bat origin. Nature 556, 255–258 (2018). https://doi.org/10.1038/s41586-018-0010-9.

[55] Y. Li et al., Potential host range of multiple SARS-like coronaviruses and an improved ACE2-Fc variant that is potent against both SARS-CoV-2 and SARS-CoV-1. bioRxiv:10.1101/2020.04.10.032342 (18 May 2020).

[56] The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification

limitation of our study is that we were currently unable to provide experimental evidence for the modification on viral RNAs." The low S/SN ratio of 1.7 in the expansion of SARS-CoV-2 in the human population would argue against a robust host RNA modification mechanism.

In summary, the findings reported here are:

1. Inconsistences between published papers and interviews as to the source and sequencing history of the original specimen that was claimed to have been collected in 2013 (RaBtCoV/4991) and the specimen for the bat RaTG13 virus. For example, two explanations of the discovery of the close relationship between RaTG13 and SARS-Cov-2, a highly homologous match between the RdRp genes of the viruses noticed in 2020 followed by full genome sequencing, or identification in 2020 of a homologous match to full genome sequencing previously done in 2018. Current publicly available data for RaTG13 from 2017 and 2018 is a set of 33 amplicon sequencing runs but they cover only about 80% of the entire genome. In the *Science* interview Dr. Shi's says the specimen for RaTG was consumed during sequencing in 2018, but if this is true, the RNA-Seq referred to in the *Nature* paper could not have been performed in 2020. At this time, the Wuhan laboratory has not met the requirements of *Nature* with respect to the sharing of primary and sequence assembly data from their seminal paper[1] and this data should be provided immediately.
2. The specimen from which RaTG13 was reported to have been isolated and which has been repeatedly reported to have been a bat fecal specimen has a taxonomical composition of eukaryotes, bacteria, and viruses that is completely different from a set of nine bat fecal specimens collected in the same field visits by the same laboratory personnel from the Wuhan Institute of Virology. The probability that an authentic fecal specimen could have the composition reported is one in ten million, an impossibly low occurrence. Examination of the strong signals in the RaTG13 specimen identifies both a variety of bat genetic material, some that are not native to China, as well as unexpected species, such as marmots and a red fox. It also contains a telltale 3% primate sequence consistent with VERO cell contamination. I propose that this specimen is apparently either a mislabeled specimen (although I cannot conjure what the field source or specimen would be) or was artificially created in a laboratory.
3. The method-dependent sequence differences between the amplicon data and the RNA-Seq data are about 5% or about 50-times higher than expected as a technical error rate of 0.1%. This is an experimental quality issue that needs to be addressed; no explanation has been offered for this to date. In addition, no assembly methodology has been provided and at least two gaps, totaling over 60 nt, were easily identified.
4. The findings, reported here of a mutational drift of synonymous mutations only between SARS-CoV-2 and RaTG13 in the Spike Protein S1/S2 region and the pp1ab gene that has never been seen in nature before and which has a probability of having occurred by chance of less than one in ten million and one in one billion makes it more likely that, at least for these portions of the RaTG13 genome, comprising over one-

> quarter of the entire genome, another process is underway. With the demonstration that codon-enhancement or optimization can produce this unnatural S/SN pattern, some form of laboratory-based synthetic biology was performed on RaTG13, SARS-CoV-2, or both.

Apparently, the entire specimen from which RaTG13 was purported to have been found has been consumed in previous sequencing experiments and the Principal Investigator has stated that no virus has ever been isolated or cultured from the specimen at any time in the past. Given the irregularities and anomalies identified in this paper it seems prudent to conclude that all data with respect to RaTG13 must be considered suspect. As such, reliance of the foundational papers of the origin of SARS-CoV-2 as having arisen from bats via a zoonotic mechanism must be reexamined and questioned.

Paper 2: The February 19, 2020 Lancet paper entitled: "Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19."

On February 19, 2020 *The Lancet* published a Correspondence entitled "Statement in support of the scientists, public health professionals, and medical professionals of China combatting COVID-19[57]" with 27 public health scientists from eight countries as authors. The statement seems to attempt to settle the question of the origin of SARS-CoV-2 and short circuit further debate, as the second sentence reads: "We stand together to strongly condemn conspiracy theories suggesting that COVID-19 does not have a natural origin." It goes on to state: "Conspiracy theories do nothing but create fear, rumors, and prejudice that jeopardize our global collaboration in the fight against this virus."

The letter provided an open solicitation for support and at this time has been signed by at over 20,300 people, as if to purport that science can be advanced through polling and the democratic process.[58] While it is a truism that conspiracy theories have no place in the academia, legitimate debate should not be foreclosed.

The statement itself provides a more nuanced discussion of the evidence for a zoonotic origin and contains 14 references, eight of which contain data about the COVID-19 pandemic and six of which are governmental policy statements without new data, background articles from 2003 and 2004 on zoonotic diseases, or a virus naming statement by the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses, which is responsible for developing the official classification of viruses and taxa naming (taxonomy) of the Coronaviridae family. The eight articles with data were written at the end of January or early February, when there were fewer than 10,000 patients.

[57] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30418-9/fulltext#back-bib1

[58] This is reminiscent of the story attributed to Albert Einstein by Stephen Hawkins in his *Brief History of Time*. According to Hawkins, a book was published in 1930 in pre-war Germany entitled, "One Hundred Authors Against Einstein." When he was asked about the book Einstein is reported to have retorted, "If I were wrong, then one would have been enough!"

An analysis of the evidence for a zoonotic source given in support of the above Statement is contained in Text-Table here. The analysis shows there was very little actual data available at the time to permit reaching such a definitive conclusion. There was also the absence of data or discussion that could support a laboratory origin.

Reference	Statements concerning origin of SARS-CoV-2	Response to statements
1.Gorbalenya AE Baker SC Baric RS et al. Severe acute respiratory syndrome-related coronavirus: the species and its viruses—a statement of the Coronavirus Study Group. bioRxiv. 2020; (published online Feb 11. DOI: 2020.02.07.937862 (preprint).)	A naming statement about SARS-CoV-2. The emergence of SARS-CoV-2 as a human pathogen in December 2019 may thus be perceived as completely independent from the SARS-CoV outbreak in 2002–2003. With respect to novelty, SARS-CoV-2 differs from the two other zoonotic coronaviruses, SARS-CoV and MERS-CoV, introduced to humans earlier in the twenty-first century.	Does not provide data on a potential zoonotic source.
2.Zhou P Yang X-L Wang X-G et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; (published online Feb 3.)	The sequences of 2019-nCoV BetaCoV/Wuhan/WIV04/2019 among patient specimens are almost identical and share 79.6% sequence identity to SARS-CoV. Furthermore, we show that 2019-nCoV is 96% identical at the whole-genome level to a bat coronavirus. Pairwise protein sequence analysis of seven conserved non-structural proteins domains show that this virus belongs to the species of SARSr-CoV. The close phylogenetic relationship to RaTG13 provides evidence that 2019-nCoV may have originated in bats.	The bat genome identity of 96% described here, coupled with the known mutation rate of SARS-CoV-2 of about 26/year, implies a **lowest common ancestor about 44 years ago.**